Video Based Reconstruction System for Mixed Reality Environments Supporting Contextualised Non-Verbal Communication and Its Study

Allen James Fairchild

UNIVERSITY OF SALFORD

School of Computing, Science and Engineering

2016

This thesis is submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Contents

List of Terms and Abbreviations vi

Abstract xvi

1 Introduction 1

1.1 Purpose 2

1.2 Motivation 2

1.3 Problem Characteristics 3

1.3.1 Spatial and Colour Calibration 3

1.3.2 Background-Foreground Segmentation 3

1.3.3 System Architecture 4

1.4 Development Methodology 4

1.5 Contributions 5

1.6 Scope 6

1.7 Thesis Overview 7

1.7.1 Background and Related Work 7

1.7.2 Spatial and Colour Calibration 7

1.7.3 Background-Foreground Segmentation 8

1.7.4 System Architecture 8

1.7.5 Discussion and Conclusion 8

1.8 Summary 8

2 Background and Related Work 10

2.1 Search Methodology 11

2.1.1 Identifying Projects 11

2.1.2 Journals 13

2.1.3 Conferences 13

2.2 Telepresence 14

2.2.1 Video Conferencing 14

2.2.2 Immersive Collaborative Virtual Environments 15

2.3 Immersive Virtuality Telepresence 16

2.4 Capturing the Three-Dimensional Form of People and Objects 16

2.5 Video Based Reconstruction Systems 18

2.5.1 Grimage 19

2.5.2 Blue-c 19

2.5.3 British Broadcasting Corporation 20

2.5.4 University of Kyoto 20

2.5.5 DreamWorld 21

2.6 Spatial and Colour Calibration 21

2.6.1 Spatial 21

2.6.2 Colour 23

2.7 Background-Foreground Segmentation 23

3 Spatial and Colour Calibration 26

3.1 Spatial 27
3.1.1 Previous Calibration Implementation 27
3.1.2 New Calibration Implementation 32
3.2 Colour 43
3.2.1 Advanced Colour Correction Method 48
3.2.2 Simple Colour Correction Method 53
3.2.3 Example Colour Correction Results 54
3.2.4 Summary 56
4 Background-Foreground Segmentation 58
4.1 Visible Light 59
4.1.1 Mester Implementation 59
4.1.2 Chroma-Keying Implementation 61
4.1.3 Summary of Previous Segmentation Implementations 62
4.1.4 Updated Segmentation Implementation 63
4.1.5 Summary 79
4.2 Infrared 80
4.2.1 Summary 94
5 System Architecture 95
5.1 System Architecture 96
5.1.1 Subject Acquisition 97
5.1.2 Capture Nodes 99

5.1.3 Capture Node Usability Enhancements 103
5.1.4 3D Model Generation 105
5.1.5 Distribution 105
5.1.6 3D Model Rendering 108
5.1.7 Proof-of-Concept End-to-End Demo 113
5.1.8 Conveying Non-Verbal Behaviour 117
6 Discussion and Conclusions 127
6.1 Discussion 128
Spatial and Colour 129
6.1.1 Calibration 129
6.1.2 Segmentation 129
6.1.3 System Architecture 130
6.2 Limitations and Future Work 130
Spatial and Colour 130
6.2.1 Calibration 130
6.2.2 Background-Foreground Segmentation 131
6.2.3 System Architecture 132
6.3 Conclusion 133
References 135
Appendix A 148
Appendix B 149

Appendix C..... 150

List of Terms and Abbreviations

2D	Two-dimensional
3D	Three-dimensional
Avatar	A Graphical Representation of a User
CCTV	Closed Circuit Television
CGI	Computer Generated Imagery
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
EPVH	Exact Polyhedral Visual Hull
FPS	Frames Per Second
GPU	Graphics Processing Unit
H.264	Advanced Video Coding (MPEG-4 AVC)
IBR	Image Based Reconstruction
ICVE	Immersive Collaborative Virtual Environment
IR	Infrared
IVT	Immersive Virtuality Telepresence
L.E.D	Light Emitting Diode
LZMA	Lempel–Ziv–Markov chain Algorithm
NVB	Non-Verbal Behaviour
NVC	Non-Verbal Communication
OpenCV	Open Computer Vision Library
OpenGL	Open Graphics Library

PC	Personal Computer
SfS	Shape-from-Silhouette
TCP/IP	Transmission Control Protocol/Internet Protocol
VBO	Vertex Buffer Objects
VBR	Video Based Reconstruction
VC	Video Conferencing
VE	Virtual Environment
VR	Virtual Reality
VRPN	Virtual Reality Peripheral Network
XML	Extensible Markup Language

List of Tables

Table 3-1 Comparison of Exposure Value to Total Images Acquired 33

Table 3-2 Comparison of Exposure Value to Total Images Acquired 40

Table 3-3 Results from Different Sphere Detection Methods (Lower Values Better)... 42

Table 4-1 Success Rate Settings..... 77

Table 4-2 Success Rates for Mester 77

Table 4-3 Success Rates for MOG 78

Table 4-4 Segmentation Times for Processing Two Streams Simultaneously..... 78

Table 5-1 XML Syntax for Camera Settings..... 103

Table 5-2 XML Syntax for the Reconstruction Node 104

List of Figures

Figure 1-1 Iterative Development Model 5

Figure 2-1 Cisco TelePresence Immersive Experience 14

Figure 2-2 ICVE displaying a CGI avatar 16

Figure 3-1 Previous Calibration Process (Blue: Automated, Orange: Manual Process) 28

Figure 3-2 Previous Wand 29

Figure 3-3 Calibrator Wearing White Suit 30

Figure 3-4 Result of Sphere Detection with Normal Exposure 31

Figure 3-5 Result of Sphere Detection with Long Exposure 31

Figure 3-6 Colour Curves for the Basler piA1000-48gc Cameras 34

Figure 3-7 Lamps and L.E.Ds used in Early Wand Development Experiments 35

Figure 3-8 Spheres of Various Size and Colour used in Early Wand Development Experiments 35

Figure 3-9 Spheres Attached to Mounting Rods 36

Figure 3-10 Illuminated Sphere Wand 37

Figure 3-11 Calibrator Holding the Illuminated Sphere Wand 38

Figure 3-12 Result of Sphere Detection with the New Method 39

Figure 3-13 New Calibration Process (Blue: Automated, Orange: Manual Process) 41

Figure 3-14 Striping Effect 44

Figure 3-15 - Professional Colour Card Positioned in Centre of Octave 46

Figure 3-16 - Luminosity Inconsistencies Between Two Cameras 47

Figure 3-17 - Syncronised Images from 10 Cameras 48

Figure 3-18 - Syncronised Images Focused on Head of Participant 49

Figure 3-19 - Brightness Adjust Taking Place 50

Figure 3-20 - Result of Brightness Correction 50

Figure 3-21 - Hue Correction Process 53

Figure 3-22 - Colour Correction Results 54

Figure 3-23 3D Avatar of User Stance A 55

Figure 3-24 3D Avatar of User Stance B 56

Figure 4-1 Poor Segmentation Result with Misclassified Shadow 60

Figure 4-2 Poor Segmentation Result with Noise 61

Figure 4-3 - 3D Model Generated Using Green Chroma-keying Method 62

Figure 4-4 - Silhouette Image Demonstrating Segmentation of Seams in Octave Floor and Detected Shadow 66

Figure 4-5 - Silhouette Image with Detected Shadow Removed 67

Figure 4-6 Final Silhouette Image with Shadows and Noise Removed 68

Figure 4-7 - 3D Hand Result using Mester (left), MOG (centre), MOG with a Single Erode Pass (right) 69

Figure 4-8 Mester (left), MOG (right) 70

Figure 4-9 Thresholded Image in Aletheia 72

Figure 4-10 Manual Removal of Components 73

Figure 4-11 Silhouette Border Idenfication 74

Figure 4-12 Silhouette Border Refinement 75

Figure 4-13 Example of Included and Excluded Areas 76

Figure 4-14 - Segmentation in Visible Light Spectrum whilst Video Projected 81

Figure 4-15 Infrared Segmentation Experimental Setup 82

Figure 4-16 Infrared Segmentation Experimental Setup 83

Figure 4-17 Infrared Segmentation Experimental Setup 84

Figure 4-18 - Infrared Segmentation Result with Moving Background..... 85
Figure 4-19 - Infrared Segmentation Result With Moving Background..... 86
Figure 4-20 Infrared Lamp Positioned Alongside Kinect Sensor 87
Figure 4-21 Infrared Lamp Positioned Slightly Above Floor Level 88
Figure 4-22 Segmentation Example with the Kinects Built-In Infrared Projector..... 88
Figure 4-23 Segmentation Example with Infrared Lamp Next to Kinect 89
Figure 4-24 Segmentation Example with Infrared Lamp Slightly Above Floor Level.. 89
Figure 4-25 Segmentation Example Using Both Infrared Lamps 90
Figure 4-26 Distance A 91
Figure 4-27 Distance B..... 91
Figure 4-28 Distance C..... 92
Figure 4-29 ThinkLab..... 93
Figure 4-30 ThinkLab Infrared Setup..... 93
Figure 4-31 ThinkLab Infrared Result..... 94
Figure 5-1 Generic System Architecture..... 96
Figure 5-2 Panoramic Image of the Octave..... 97
Figure 5-3 UML Deployment Diagram of the Octave 98
Figure 5-4 End-to-End System UML Deployment Diagram 99
Figure 5-5 Image Acquisition and Reconstruction Threads and Buffer..... 101
Figure 5-6 Calibration Message Format..... 106
Figure 5-7 Network Message Format..... 107
Figure 5-8 Video Stream Message Format..... 107
Figure 5-9 Time Management and Rendering Threads..... 109
Figure 5-10 Texturing Process 111

Figure 5-11 Texturing Process 113
Figure 5-12 Demo Linkup Architecture 115
Figure 5-13 Proof-of-Concept Example One 116
Figure 5-14 Proof-of-Concept Example Two 116
Figure 5-15 Proof-of-Concept Example Three 117
Figure 5-16 Proof-of-Concept Example Four 117
Figure 5-17 User Pointing 118
Figure 5-18 Two 3D Reconstructed Users and CGI Avatar Waving 119
Figure 5-19 Demonstrating Interpersonal Distance Between 3D Avatar and CGI Avatar 119
Figure 5-20 Universal facial expressions of emotion and eye gaze. This figure shows a 3D reconstruction of the author attempting to recreate the seven universal emotions and eye gaze 120
Figure 5-21 User Holding and Writing on a Chalkboard View A 120
Figure 5-22 User Holding and Writing on a Chalkboard View B 121
Figure 5-23 User Holding and Writing on a Chalkboard View C 121
Figure 5-24 Diagram of Setup of the Asymmetric Telepresence System 122
Figure 5-25 Client Looking at an Approaching Virtual Threat that is Prerecorded 123
Figure 5-26 The Therapist (Captured Live) Moves Between Client and Threat, and Tries to Redirect Client Attention (View A) 123
Figure 5-27 The Therapist (Captured Live) Moves Between Client and Threat, and Tries to Redirect Client Attention (View B) 124
Figure 5-28 User Being Acquired in the Octave 124

Figure 5-29 3D Avatar of User Projected onto Chromatte Material in Remote Location (Position A) 125

Figure 5-30 3D Avatar of User Projected onto Chromatte Material in Remote Location (Position B)..... 126

Figure 6-1 Example of Infrared Lamp and Camera Configuration..... 132

Acknowledgements

I would like to take this opportunity to express thanks to my supervisor, David Roberts, for his support, guidance and enthusiasm.

To Carl Moore and Tobias Duckworth, for their research and the generous and supportive manner in which they transferred it to me.

To Robin Wolff, Simon Campion and Arturo Garcia, it has been a pleasure working with such a dedicated group of researchers in a collaborative project where our passionate commitment has enriched the quality of research.

Thanks to John O'Hare for all the time he has spent with me reconfiguring the Octave, I wish him all the best with his own research.

Thanks to Apostolos Antonacopoulos, Christian Clausner, Christos Papadopoulos and Stefan Pletschacher for their assistance and guidance. It was valuable to talk to these talented individuals.

Thanks to my mother, father, family and friends for their support and encouragement as I embarked on this arduous journey.

Last and certainly not least I would like to take this opportunity to thank my wife Nichola for her support, encouragement and patience and my daughter, Jessica, who has given me motivation and made everything that I do all the more worthwhile.

Declaration

Parts of this work have been peer reviewed and published in the following outlets:

IEEE Transactions on Circuits and Systems for Video Technology, A Mixed Reality Telepresence System for Collaborative Space Operation, Accepted.

International Conference on Disability, Virtual Reality & Associated Technologies (ICDVRAT), Bringing the client and therapist together in Virtual Reality Telepresence Exposure Therapy, Accepted.

IEEE Journal of Selected Topics in Signal Processing, withyou—An Experimental End-to-End Telepresence System Using Video-Based Reconstruction, Accepted.

IEEE Virtual Reality, Collaborative Telepresence Workspaces for Space Operation and Science, Accepted.

Abstract

This Thesis presents a system to capture, reconstruct and render the three-dimensional form of people and objects of interest in such detail that the spatial and visual aspects of non-verbal behaviour can be communicated.

The system supports live distribution and simultaneous rendering in multiple locations enabling the apparent teleportation of people and objects. Additionally, the system allows for the recording of live sessions and their playback in natural time with free-viewpoint.

It utilises components of a video based reconstruction and a distributed video implementation to create an end-to-end system that can operate in real-time and on commodity hardware.

The research addresses the specific challenges of spatial and colour calibration, segmentation and overall system architecture to overcome technical barriers, the requirement of domain specific knowledge to setup and generate avatars to a consistent high quality.

Applications of the system include, but are not limited to, telepresence, where the computer generated avatars used in Immersive Collaborative Virtual Environments can be replaced with ones that are faithful of the people they represent and supporting researchers in their study of human communication such as gaze, inter-personal distance and facial expression.

The system has been adopted in other research projects and is integrated with a mixed reality application where, during a live linkup, a three-dimensional avatar is streamed to multiple end-points across different countries.

Chapter 1

Introduction

The introduction begins by presenting the purpose of the thesis and then discusses the motivation behind it. Next, the problem characteristics identified are discussed followed by the methodology adopted for development. The introduction concludes by detailing the contributions that have been made and presenting a short chapter summary.

1.1 Purpose

This thesis presents a complete system for capturing, reconstructing and rendering the three-dimensional form of humans and objects. The reconstructed form can be simultaneously distributed to - and rendered in - multiple locations. If desired, the form may also be recorded and played back in natural time with free-viewpoint update. Emphasis has been placed on making the system (and subsequently immersive free-viewpoint video) useable for a wide range of research. Additionally, the thesis presents a method to capture in everyday environments, mixed display environments and a mixture of both.

1.2 Motivation

Capturing, reconstructing and rendering the three-dimensional form of people and objects has many applications and been the focus of much academic research. The motivation for the thesis is the desire for a useable system that can reproduce a wide range of non-verbal behaviour using commodity hardware and software synchronised cameras.

To create the system described in this thesis the research analysed then utilised both a parallelised version of a 3D visual hull reconstruction algorithm (Tobias Duckworth & Roberts, 2014) and a distributed video system (Moore, 2012). This was necessary in order to evaluate previous efforts at combining these systems as there had not been an attempt at creating a useable integrated system.

Developing a useable end-to-end system is important not only because it can be used by others for their research but also because it highlights issues that can be missed when solely concentrating on specific discrete problems. A complete system can help to create better methods, provide understanding and validate methods in real environments.

1.3 Problem Characteristics

The evaluation of the previous research prototypes highlighted three main areas that required addressing: spatial and colour calibration, foreground-background segmentation and system architecture. Decisions were taken about which components of the prototypes could be retained and developed further and those that didn't warrant inclusion. Then, both the reconstruction and distributed video system, were heavily modified and extended to improve performance and provide better integration in the overall architecture of the system developed.

1.3.1 Spatial and Colour Calibration

The spatial calibration implementation previously employed frequently produced sub-optimal results. The sub-optimal results produced unfaithful reconstructions and required the calibration process to be repeated. This was a time consuming process, limited the usability of system and dissuaded users from changing camera pose. The difference in colour across the cameras used to acquire the subjects also introduced issues with the faithfulness of generated avatar.

1.3.2 Background-Foreground Segmentation

The background-foreground segmentation methods employed required a tedious and time consuming setup procedure and even when configured correctly, the resultant silhouettes often contained artefacts that hindered the VBR process. To aid the segmentation process the cameras were configured with different exposure and colour settings dependant on their pose, which lead to inconsistent colour across the avatars form. Furthermore, the segmentation methods only function in sterile environments, such as a CAVE with the

monitor walls set to a consistent colour or off, so it was not possible to fully immerse a subject into a natural or dynamic environment.

1.3.3 System Architecture

The system architecture of the previous research prototypes supported a single location to render the 3D avatar and did not scale. The design required that the reconstruction process be executed on the same computer as the rendering. This is a fundamental problem for two reasons:

1. It puts an excessive hardware requirement on the location in which the rendering is taking place
2. It hinders the possibility of rendering in more than a single location

Furthermore, recording and playback was achieved by storing images to disk and loading them back into the reconstruction process without the temporal information required to allow natural playback. Furthermore, saving to disk was not supported whilst simultaneously performing a live reconstruction.

1.4 Development Methodology

Prior to system development a literature review was conducted to gain a better overall understanding of the field. Following this, a comprehensive analysis of the local research groups' prototypes and publications was conducted to evaluate then determine how best to proceed and in which specific areas to concentrate efforts. Furthermore, some general requirements from typical users of the system were elicited and then further literature searches conducted to determine if another overall approach be worthwhile considering or whether if individual components could be substituted.

Next, development methodologies where researched to ascertain those that would suit the requirements of developing a research platform. The model needed to accommodate subtle improvements or complete changes in direction post component testing and experimentation. The Iterative Development Model (Larman & Basili, 2003) sets time aside to revise and improve parts of the system (Cockburn, 2008; Incremental (2008)), thus, fitted the requirements. Furthermore, it was suited to the authors style of development. The model is shown in Figure 1-1.

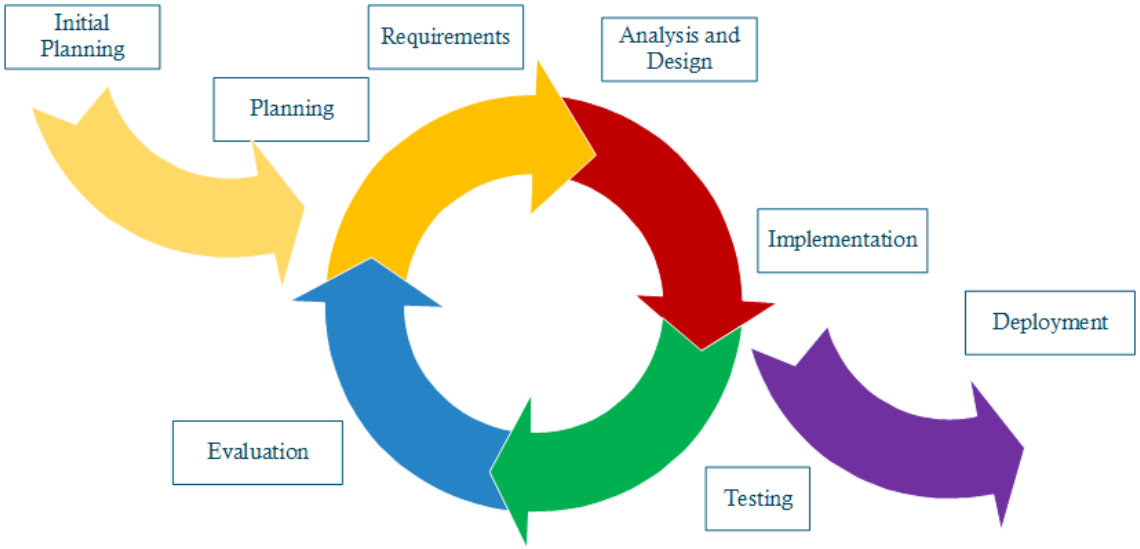


Figure 1-1 Iterative Development Model

After the initial development plan was drafted the subsequent stages of implementation and testing were executed. At the end of each phase the results were evaluated. The feedback from evaluation fed back in to the development cycle steering its course.

1.5 Contributions

The main contribution of this research is a complete video based reconstruction system for mixed reality environments with specific contributions as follows:

- A useable system architecture that can be deployed on commodity hardware and supports simultaneous rendering in multiple locations that can be geographically dispersed
- The ability to record and playback sessions in natural time with free-viewpoint update enabling researchers to study non-verbal behaviour post capture
- Straightforward and accurate calibration to enable researchers to experiment with different cameras poses
- Straightforward and robust segmentation that results in consistently faithful avatars
- A method to acquire subjects in mixed environments where the background could be both static and/or dynamic
- Methods to evaluate the background-foreground segmentation results
- Integration in to an application demonstrating proof-of-concept with faithful 3D avatars

1.6 Scope

The research presents a complete Video Based Reconstruction system, which could be relevant to entertainment, training and social psychology but is considered here only in the context of telepresence.

It presents a system to capture and faithfully recreate non-verbal behaviour but does not perform a study of it.

The resulting system should support rendering of the reconstruction to a visual quality that allows the whites of the eyes to be seen. The temporal quality of the system should be sufficient to support the conveyance of non-verbal behaviour.

The system development will focus on utilising equipment present in the Salford University Octave Research Facility (O'Hare) and other commodity hardware currently available.

1.7 Thesis Overview

The remainder of this thesis is organised as follows:

1.7.1 Background and Related Work

The background begins by detailing the search methodology that has been employed throughout the research process. Then it provides the reader with an overview of telepresence, which is important as it helps frame the work in the context of the research group's goals. Then it provides details of other telepresence systems that have been developed in the academic community. Following on, it reviews different methods that can be used to generate three-dimensional avatars of people and objects providing justification for continuing with a video based reconstruction approach. Then methods to calibrate both spatial and colour are reviewed. Finally, methods to perform background-foreground segmentation are explored.

1.7.2 Spatial and Colour Calibration

In this chapter the spatial and colour calibration implementations of the system are detailed. The accuracy of spatial calibration is directly linked to the quality of the models generated and the consistency of the colour across the cameras has an impact on their faithfulness and believability. Both need to be accurate to realise consistently faithful reconstructions. First, the spatial calibration implementation is described, which has automated most steps of the calibration (resulting in a faster process), reduced the re-projection error (thus improved quality) and the removed the requirement of wearing

white clothes. Next, the colour calibration implementations, which have produced greater homogeneity among cameras are discussed.

1.7.3 Background-Foreground Segmentation

This chapter presents the background-foreground segmentation improvements that have been contributed to the system. It begins by analysing the current methods employed in the prototype to gain an understanding of the challenges and current shortcomings. Then it presents the steps taken to improve results in visible light and finally demonstrates an approach to perform segmentation in the Infrared light spectrum.

1.7.4 System Architecture

In this chapter the end-to-end system architecture is presented. The process begins with the acquisition then segmentation of the subject(s) and object(s) of interest, followed by the 3D reconstruction, then the distribution and rendering. First, the end-to-end system architecture is detailed followed by a description of each stage in the pipeline.

1.7.5 Discussion and Conclusion

The thesis concludes with a discussion about the system presented and the research that was undertaken during its development. The purpose of the system, motivation for developing, problem characteristics identified and the contributions made are revisited. Finally, the limitations of the system and research are identified, conclusions drawn and possible future development suggested.

1.8 Summary

The introduction began by presenting the purpose of the thesis, and then discussed the motivation behind it. It continued with identifying and discussing problem characteristics

before outlining the methodology adopted for development. Next, the contributions of the research were outlined and scope of the work discussed. Finally, it concluded with a thesis overview.

Chapter 2

Background and Related Work

The background begins by detailing the search methodology that has been employed throughout the research process. Then it provides the reader with an overview of telepresence, which is important as it helps frame the work in the context of the research group's goals. Next, it provides details of other telepresence system that have been developed in the academic community. Following on, it reviews different methods that can be used to generate three-dimensional avatars of people and objects providing justification for continuing with a video based reconstruction approach. Then methods to calibrate both spatial and colour are reviewed. Finally, methods to perform background-foreground segmentation are explored.

2.1 Search Methodology

The search methodology describes how relevant literature has been acquired, thus enabling the searches to be repeated for the discovery of new literature at a later date. The method is discussed in the following subsections:

2.1.1 Identifying Projects

The first step was identifying similar projects. This was achieved by conversing with fellow academics who have extensive knowledge of the field and running searches in Google Scholar with index terms as follows:

3D telepresence, 3D video reconstruction, immersive telepresence

The projects identified along with a brief description are detailed in the following subsections.

2.1.1.1 BEAMING

The EU collaboration BEAMING project aims to instantaneously virtually transport people (visitors) from one physical place in the world to another (the destination) so that they can interact with the local people there, allowing for applications such as collaboration, gaming, entertainment and teleoperation ("BEAMING : Being in Augmented Multi-Modal Naturally-Networked Gatherings," 2013).

2.1.1.2 blue-c

The blue-c project combines the qualities of total immersion experienced in CAVE™-like environments with simultaneous, real-time 3D video acquisition and rendering from multiple cameras to enable a number of participants to interact and collaborate inside an

immersive, virtual world, while perceiving the photorealistic three-dimensional human inlays of their collaboration partners in real time ("blue-c," 2003).

2.1.1.3 Blue-c II

The aim of the blue-c II project is closely linked to the desired impact this research will have in that its focus is to investigate and develop fundamental methods for collaboration environments and multi-modal acquisition and interaction with large and complex physical environments comprising interactive, view-independent 2D and 3D video, display technology, and natural and non-intrusive interaction. It will see cameras and projectors being integrated in large scenes or in productive, office-like environments ("blue-c-II," 2012).

2.1.1.4 Office of the Future

The Office of the Future project at the University of North Carolina is based upon a unified application of computer vision and computer graphics in a system that combines and builds upon the panoramic image display, tiled display systems, image-based modelling, and immersive environments the goals of which include realising an everyday graphical display environment, and 3D tele-immersion capabilities allowing distant people to feel as though they are together in a shared office space ("Office of the Future," 2009). The projects goal of using computer vision techniques, in real time, to dynamically extract per-pixel depth and reflectance information for the visible surfaces in the office, such as walls, furniture, objects and people is closely related to the intended focus of this research.

After the projects were identified, contributing authors and publication lists were acquired. The authors identified were then searched for on ResearchGate which led to

other relevant experts working in the field to be discovered. Authors of significant interest were: Henry Fuchs, Adrian Hilton and Oliver Grau (and their research groups or corporate organisations they belong to).

IEEE Xplore, Science Direct, Google Scholar and the BBC Research and Development repository have been used to search for publications using both keywords, paper titles and authors names. Then CiteSeerX has been utilised to determine others publications of interest that have been referenced in those identified. The databases that have been identified as applicable to this field and my discipline are IEEE and ACM. The following journals and conferences have also been identified:

2.1.2 Journals

- MIT Presence: Teleoperators and Virtual Environments (Press)
- Springer Journal of Real-Time Image Processing (Springer)
- IEEE Transactions on Visualization and Computer Graphics (Society)
- IEEE Transactions on Multimedia (IEEE)

2.1.3 Conferences

- ACM Multimedia (SIGMM)
- Computer Supported Cooperative Work and Social Computing (ACM)
- CHI Conference on Human Factors in Computing System (CHI)
- Special Interest Group on Computer Graphics and Interactive Techniques (Machinery)
- IEEE VR (IEEE)
- 3DTV Conference (IEEE)

These journals and conferences and research groups will be periodically reviewed for updated information and related work and where applicable they will also be published to.

2.2 Telepresence

Telepresence is a term used to describe technologies that attempt to reproduce face-to-face meetings where people appear present in a location that is remote (and often geographically dispersed) from their own.

2.2.1 Video Conferencing

Video conferencing (VC) can be achieved using free software such as Skype or more advanced commercial offerings such as Cisco TelePresence (Figure 2-1). Appearance is faithfully transmitted via VC and careful alignment of camera and screen can create the illusion of eye contact as long as the participants are in fixed positions and do not move, but gaze-direction and focus of attention are not evident.

Figure 2-1 Cisco TelePresence Immersive Experience

2.2.2 Immersive Collaborative Virtual Environments

Immersive Collaborative Virtual Environments (ICVE) featuring Immersive Projection Technology allow users to be immersed within a multitude of environments not limited to say a fire and rescue scenario (Backlund et al., 2007), battlefield (Isabelle et al., 1997) or as is more applicable to telepresence: an office or living room. ICVE systems typically consist of spaces where virtual content is projected on to the surrounding walls and feature technologies such as cameras, depth sensors and motion tracking systems to track and acquire images and video of those within. They can also contain systems to convey information to the participants: projection systems to convey 2D and 3D imagery (the latter often requiring users to wear 3D shutter glasses), wave field synthesis audio simulation systems and, if required, haptic feedback systems such as the phantom (Massie & Salisbury, 1994) can be present. In ICVEs remote participants have typically been represented by a traditional avatar, which may be tailored to an individual offline so that it resembles them. The avatars are brought to life by following the motion-tracked movements of the person they mimic but seldom convey facial expression, body torque or finger gestures, all of which are important parts of human communication because numerous markers must be painstakingly placed, which is impractical for everyday meetings. Figure 2-2 shows an example of a simple CGI avatar projected on to a monitor wall of an ICVE.

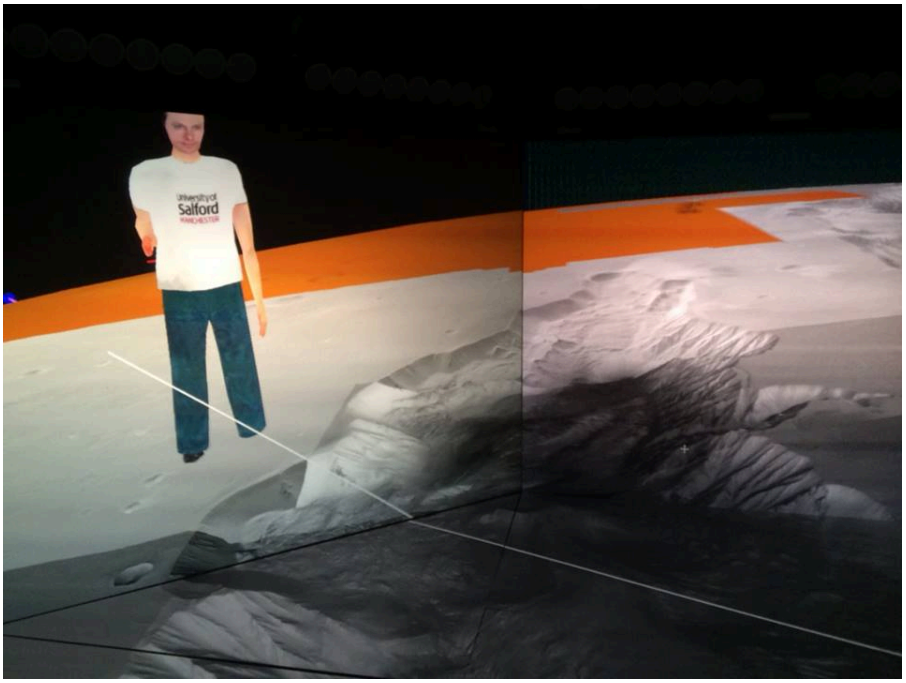


Figure 2-2 ICVE displaying a CGI avatar

2.3 Immersive Virtuality Telepresence

VC can faithfully convey what participants look like, whilst ICVEs can faithfully convey what participants are looking at, but neither can achieve both. It is desirable to achieve immersive collaboration using realistic 3D avatars and this work builds upon previous research seeking to do that. Immersive Virtuality Telepresence (IVT) is a term used to describe the creation of realistic virtual avatars of people that can be projected into immersive environments to provide the illusion that they are actually present (Tobias Duckworth & Roberts, 2014).

2.4 Capturing the Three-Dimensional Form of People and Objects

There are a number of approaches to generate 3D avatars of users and they fall into two main categories: active and passive. Active methods include time-of-flight devices that project light towards and analyse the time it takes to reach points on an object (Hansard et al., 2012), and structured light devices that analyse disparity in a projected pattern to

form a 3D representation (Stockman et al., 1988; Will & Pennington, 1971). The Kinect is an example of a structure light device that has been used in telepresence research. A single Kinect can achieve a partial 3D reconstruction of the subject in the plane it is pointing towards. However, we need to generate a complete 3D reconstruction of the user in order to place him/her inside the Virtual Environment and allow them to be viewed from any angle, and that would mean the use of multiple Kinects to be positioned around the subject. Herein lies a problem, because the projected patterns from the individual Kinects interfere with one another, and this causes deterioration in the quality of the depths maps, typically resulting in less faithful shapes with holes in them. It has been demonstrated that the interference between multiple Kinects can be reduced (Maimone & Fuchs, 2012b) and there are numerous examples of Kinects being used for 3D capture (Maimone & Fuchs, 2011a, 2011b, 2012a) but to the authors knowledge only one where a 3D avatar is generated without the surrounding environment (Alexiadis et al., 2013).

With passive methods, also known as Image Based Reconstruction (IBR) techniques (Debevec et al., 1996), natural images such as those acquired from a conventional camera capturing light in the visible spectrum are used to construct a 3D model of an object by using more than one viewpoint. Video Based Reconstruction (VBR), which can be achieved by extending IBR into four dimensions, is a technique used to create dynamic three-dimensional models from video streams. There are several VBR approaches suitable for reconstructing the 3D form of an entire human, a key requirement for our system. One example is multi-view stereo (Furukawa & Ponce, 2010), which is capable of producing high quality, spatially accurate and visually faithful models. Unfortunately, it currently falls short of the temporal requirements of a real time telepresence system. Techniques based on the shape-from-silhouette (SfS) principle (Baumgart, 1975), which

form an approximation to the 3D shape known as the visual hull(Laurentini, 1994), has demonstrated that it can fulfil this requirement whilst retaining a faithful reconstruction (Roberts et al., 2015).

Both active and passive methods have strengths and weaknesses when applied to 3D telepresence avatar generation. Multiple Kinect based approaches are currently of a lower resolution compared to that which can be achieved with SfS using conventional cameras with resolutions typically in excess of 1000x1000 compared to 320x240 pixels depth map resolution. They offer a less faithful reproduction because the holes produced due to pattern interference need to be filled and what fills them may not be a true representation of the real world. Moreover, there is a drop in quality of depth maps over distance (Tong et al., 2012) and this reduces the potential capture volume, which is not desirable for user movement or interacting with objects. SfS currently offers higher textural resolution and does not suffer from holes thus enabling clearer representation of eye gaze and facial expressions both of which are vital for portraying accurate NVB. Depth based approaches, however, can be deployed within an immersive environment where as, with the exception of (Lee et al., 2004), SfS requires a sterile background that would prevent the system to fully immerse the user.

2.5 Video Based Reconstruction Systems

The previous subsection explored various methods capable of capturing the 3D form with an emphasis being placed on those applicable to real-time systems. This subsection presents systems using Video Based Reconstruction and more specifically the Shape-from-Silhouette approach.

2.5.1 Grimage

The team at INRIA developed a multi camera real-time 3D modelling system for telepresence and remote collaboration applications named Grimage (Petit et al., 2009). It is capable of processing the input from up to 16 FireWire one-megapixel resolution cameras. Each camera is connected to its own dedicated computer that performs image processing. To minimise network bandwidth requirements, only the portion of the colour image required to texture the reconstructed model is sent thus reducing the amount of data by a ratio of approximately 5:1. Subsequent sub-sampling and compression increases the ratio to 20:1. Sending the sub-sampled silhouette data along with the colour image is required to decode it. The combined compression ratio for both silhouette and colour video data being 16.55:1. The bandwidth requirement for the eight cameras acquiring frames at 20 fps is 232Mbit/s, which as the authors suggest, does leave room for scaling on a Gigabit network (1000Mbit/s). However, the figure is calculated on the assumption that a participant is only occupying 20% of the image. If this were to increase (for example: if multiple participants were present in the capture volume) then it could in fact exceed the bandwidth constraints. It uses a parallel processing cluster comprising of 10 computers to calculate the visual hull.

2.5.2 Blue-c

Blue-c (Gross et al., 2003) is a sophisticated system utilising custom projection screens that can be switched from a whitish opaque state (for projection) to a transparent state (for acquisition), which allows the video cameras to “look through” the walls and perform segmentation. It has substantial requirements such as: purpose built hardware, clusters of PCs for parallel processing requirements and accurate temporal synchronisation of camera shutters, L.E.D illumination devices and screen opacity. Gigabit Ethernet or ATM

OC3 is the network transport medium suggesting that dedicated leased lines may be required to support rendering at geographically remote sites. It achieves a reconstruction rate of nine frames per second for a 25k point cloud and five for containing 15k points.

2.5.3 British Broadcasting Corporation

The BBC have used VBR approaches to provide actor feedback in computer generated environments (Grau et al., 2003; Grau et al., 2004) where a retroreflective material, typically Chromatte (Reflecmedia, 2015), is used to allow projection from one angle and simultaneous segmentation of an actor from another. In addition, they also developed a prototype system that enables free-viewpoint video of sport scenes (Grau et al., 2007). Both systems use hardware synchronisation throughout. The former system processes in real time whereas the latter performs the 3D reconstruction offline.

2.5.4 University of Kyoto

The University of Kyoto have developed two Video Based Reconstruction systems and although the systems were developed for digital 3D archiving and TV production and not real-time systems they do merit attention. The first system uses 25 VGA (640x480) cameras and a parallel computation cluster of 30 computers to calculate the 3D form (Matsuyama et al., 2004). The second system utilises 15 XGA cameras with each connected to a dedicated PC capture node (Starck et al., 2009) and a single PC (in an offline process) to perform the reconstruction process which took on average of a minute per frame on an Intel(R) Xeon(TM) 3.6 GHz CPU and 38 minutes per frame on a an Intel(R) Xeon(TM) 3GHz CPU.

2.5.5 DreamWorld

DreamWorld (Shujun et al., 2009) uses a GPU accelerated approach and is capable of processing the input from six VGA resolution cameras at 20 frames per second. To facilitate compression only the portion of the image contained within the silhouette is sent to the reconstruction server.

2.6 Spatial and Colour Calibration

2.6.1 Spatial

The faithfulness of reconstructed avatars is dependent on the accuracy of spatial calibration. The cameras are required to know where each one is relative to one another in real world space. The process of camera calibration involves determining the relationship between 3D world coordinates and the camera image plane (Tsai, 1987). Complete calibration requires determining both intrinsic camera parameters (focal length and lens distortion) and the extrinsic parameters (location of the cameras and their orientation).

2.6.1.1 Narrowband

In applications with a few cameras positioned in a manner where they can simultaneously view an 2D plane, e.g. a stereoscopic setup, a calibration chart of known configuration can be used to calibrate the intrinsic and extrinsic parameters. The chart is positioned in the field of view of all the cameras so that they can simultaneously establish points where the cameras images correspond (Heikkila & Silvén, 1997; Tsai, 1987; Z. Zhang, 2000).

Although suitable for the calibration of a pair of cameras for stereo vision (narrowband) it is not feasible to use it for calibrating a multi camera setup required for 3D avatar generation where the cameras are dispersed (wideband) as it is difficult for all cameras to

view all of the points simultaneously. Alternative methods as discussed in the following subsection.

2.6.1.2 Wideband

For wideband, where the cameras are positioned a variety of methods have been developed, including but not limited to, using:

- A one dimensional calibration object with at least three collinear points with known relative positioning to one where the end of the line is fixed (Z. Zhang, 2004),
- A multi-view stereo system and matching features in many views then using bundle adjustment to derive camera calibration (Furukawa & Ponce, 2008),
- Images containing at least three spheres and grid pattern (H. Zhang et al., 2007),
- A single image of a globe to calibrate multiple cameras where both intrinsic and extrinsic parameters are determined (Shen et al., 2008),
- Silhouette information to calibrate multiple camera calibration based on error function derived from mutual consistency of silhouettes in pairs of views (Ramanathan et al., 2000),
- A sequence of silhouettes from images under circular motion (Huang & Lai, 2008),
- A GPU hardware accelerated approach similar to the above (Shu et al., 2008),
- A repeat hypothesis and verification process to gradually refine the calibration and synchronisation of unsynchronised cameras on a network (Sinha & Pollefeys, 2004),
- Only silhouette information in video streams (Pollefeys et al., 2009),

- A wand based calibration technique (Mitchelson & Hilton, 2003).

2.6.2 Colour

Ideally, all cameras present in a multi camera setup would display exactly the same spectral response and changes to basic colour parameters, such as brightness, contrast, exposure and gain would affect the output identically. However, this ideal situation is difficult to achieve using commodity hardware as it requires identical configuration of cameras and sensors and lenses from the same manufacturing batch. Research has been conducted to correct the differences in colour across cameras: Ilie & Welch (2005) presented a photometric calibration approach that tries to adjust settings of several cameras so that their response functions are matched as closely as possible. Joshi (2004) presented a method of calibrating an array of image sensors using a colour target of known luminance and Porikli (2003) presented a solution to the inter-camera colour calibration problem which does not require a calibration chart.

2.7 Background-Foreground Segmentation

Background-foreground segmentation is the process of separating the foreground from the background in a sequence of moving images. It has several application domains including but not limited to:

- Television and film production,
- Special effects production,
- Video content analysis in CCTV applications,
- Video based reconstruction

The result of the process is typically a binary image containing what is classified as foreground and background. The foreground is commonly referred to as the silhouette of

the object(s) intent on being reconstructed. It is required in video based reconstruction systems to calculate what is referred to as the visual hull which is the 3D volume contained within the intersecting silhouettes (Laurentini, 1994; Matusik et al., 2001; Matusik et al., 2000).

There are numerous approaches to performing background segmentation, the simplest being taking a reference frame and then looking for differences in later frames classing them as foreground pixels (Benezeth et al., 2008). Another approach is that of Chroma-keying (Schultz, 2006; van den Bergh & Lalioti, 1999) where pixels matching the predefined colour of a background material are deleted. Chroma-keying is typically used in film and television production as the method of integrating actors or presenters with computer generated imagery. It is an established approach with dedicated hardware available to perform the process.

An statistical illumination-invariant change detection method (Mester et al., 2001) has been used in other 3D VBR systems and a similar GPU accelerated implementation (Griesser et al., 2005) has been regarded for its real-time performance characteristic.

Research in the field of real-time object tracking in CCTV video (Grimson et al., 1998; Stauffer & Grimson, 1999, 2000) has contributed methods of segmentation but their primary goal was not completely faithful silhouette creation. To assist with the tracking process more advance methods of segmentation, many Gaussian Mixture-based and some featuring shadow detection, have been developed (Godbehere et al., 2012; KaewTraKulPong & Bowden, 2002; Li et al., 2003; Zivkovic, 2004; Zivkovic & van der Heijden, 2006).

Depth based segmentation approaches such as the real-time technique presented in (Abramov et al., 2012) can segment objects from the background.

Thermal keying methods such as Thermo-key (K. Yasuda et al., 2003; K. N. Yasuda, T. ; Harashima, H. , 2004) can segment users though they do require relatively expensive thermal imaging cameras and complex setup.

Chapter 3

Spatial and Colour Calibration

In this chapter the spatial and colour calibration of the system is detailed. The accuracy of spatial calibration is directly linked to the quality of the models generated by the reconstruction process. The consistency of the colour across the cameras also has an impact on the faithfulness and believability of the avatars. Both need to be accurate to realise consistently faithful reconstructions. It begins by addressing the spatial calibration and then continues with colour.

3.1 Spatial

Spatial camera calibration is the process of determining how real world 3D coordinates map to the 2D camera image plane (Tsai, 1987) and during the literature review a number of approaches to achieve it were identified. One such method, a wand based calibration technique (Mitchelson & Hilton, 2003), demonstrated particular promise as it was claimed to be robust and didn't require expensive or complicated apparatus. The method was in keeping with the desired aims of this system and had previously been adopted to calibrate the proof-of-concept end-to-end video-based 3D reconstruction system. However, in this implementation the process was lengthy and required many manual steps that could only be reproduced by those with specific knowledge of the process. Furthermore, it often produced results that were invalid or sub-optimal, thus requiring the whole process to be repeated. This resulted in a lengthy calibration procedure and deterred experimentation with different camera poses.

Before proceeding to implement the wand calibration method it was necessary to understand the shortcomings of the previous implementation so it is discussed in the following subsection.

3.1.1 Previous Calibration Implementation

The complete calibration process involves a number of stages that are visualised in Figure 3-1 below and detailed in the following section.

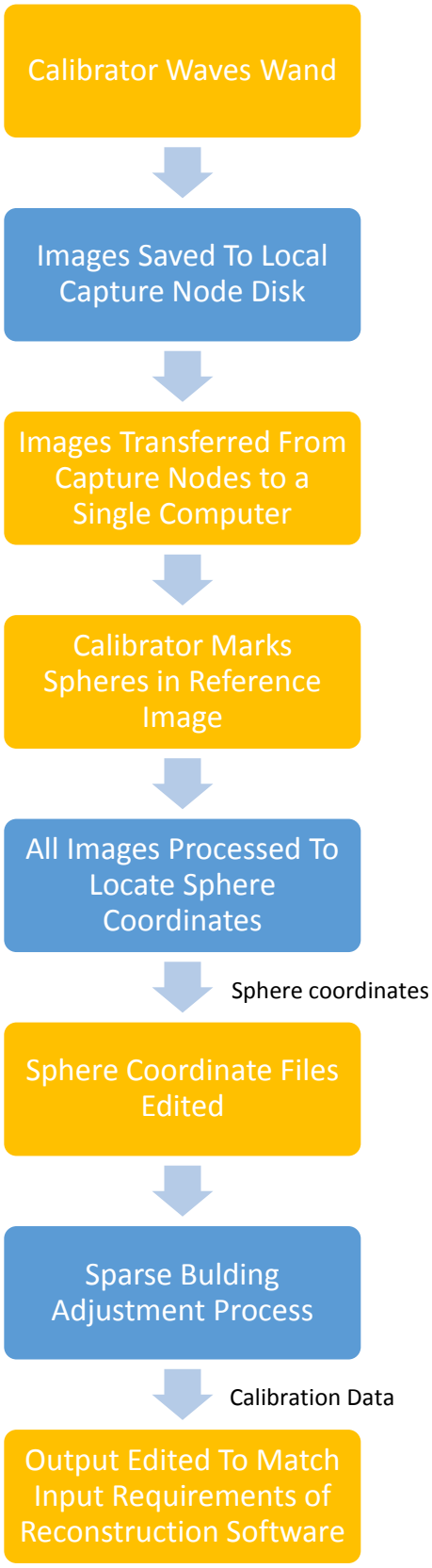


Figure 3-1 Previous Calibration Process (Blue: Automated, Orange: Manual Process)

3.1.1.1 Sphere Coordinate Extraction

The first stage of calibration is to locate the sphere coordinates simultaneously across all cameras as the wand is waved throughout the capture volume. This stage is split into two steps:

The first step in acquiring the coordinates of the spheres is to wave a wand with two different coloured spheres of a known separation mounted on it (Figure 3-2) throughout the capture volume.

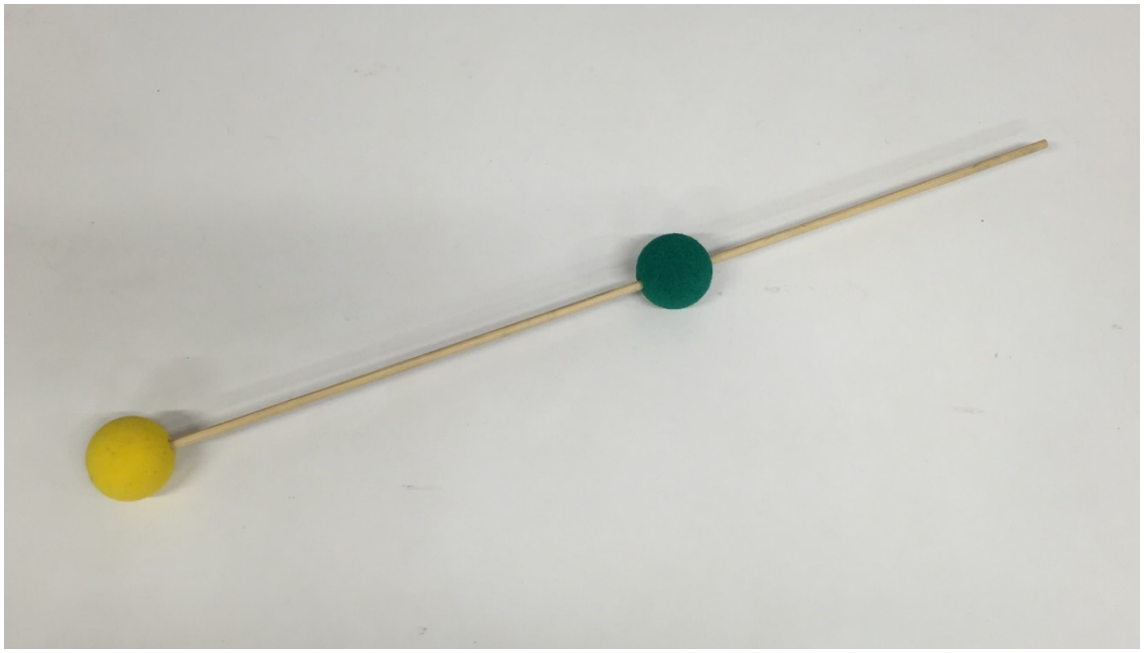


Figure 3-2 Previous Wand

The calibrator wears a white overall suit (Figure 3-3) to ensure that the scene remains as sterile as possible.

Figure 3-3 Calibrator Wearing White Suit

The resultant images are saved to disk on each of the capture nodes.

In the second step the images are copied from the capture nodes to a PC for further processing using custom software to extract the sphere coordinates. The process of copying the images is a time consuming process as there are thousands of images amounting to several gigabytes of data.

The process of extracting the sphere coordinates from the images with the custom software begins with a user marking the centre of each sphere in a reference frame. The marking records the colour of each sphere so that the colour and thus the sphere can be detected in subsequent images.

All the images are thresholded for the colours previously recorded and the resultant binary representation is then processed to acquire the sphere coordinates. The sphere coordinates are determined by placing a bounding box around the connected component in the image and calculating its centre.

In an attempt to acquire as many sphere coordinates as possible the cameras were initially configured with a normal exposure value to be capable of capturing higher frame rates. Unfortunately, this led to poor detection rates especially for the green sphere that without scene illumination had a hue that was too similar to the background.

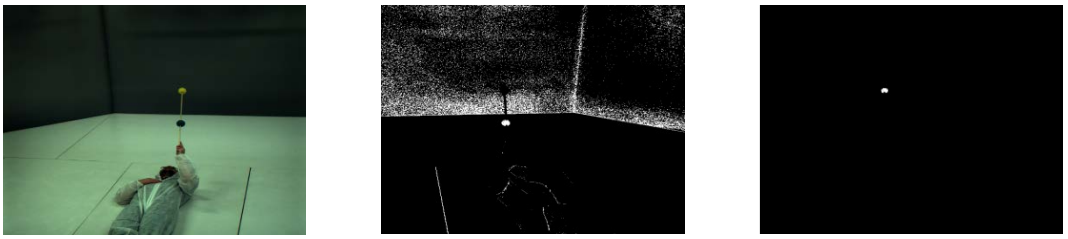


Figure 3-4 Result of Sphere Detection with Normal Exposure

In order to successfully locate the balls via there colour with the it was determined by experimentation that all room lights must be on, all projectors set to white and cameras configured with a long exposure and high gain. This produces far better results for sphere detection as shown in Figure 3-5:

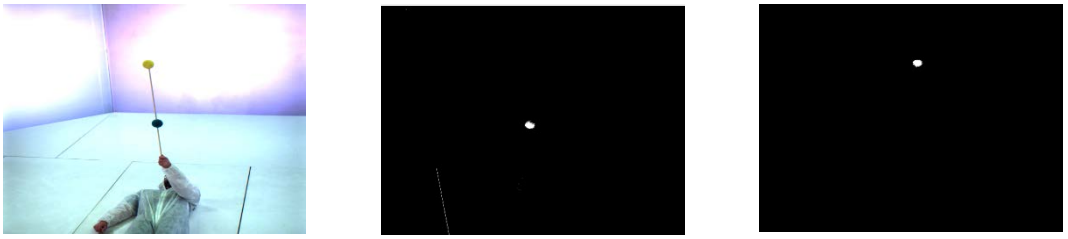


Figure 3-5 Result of Sphere Detection with Long Exposure

Yet still the balls are not exactly the correct shape and the high exposure results in fewer frames being acquired than would be achievable with a typical camera exposure.

After the sphere coordinates have been extracted and written to disk they manually edited to include the coordinate count on the first line ready to be processed in the next stage.

3.1.1.2 Sparse Bundle Adjustment

The multi-camera wand calibration software is executed and parses the coordinate files for each camera. The resultant calibration data is then manually edited to fit the requirements of the 3D reconstruction software.

3.1.1.3 Summary

The method was claimed to be robust provided the colours present in the scene were different from the colours chosen for the balls, however in practice, experimentation highlighted that in previous implementation the method of locating the centroid of the coloured balls successfully in each frame proved problematic.

Frequent repeat attempts to acquire enough valid sphere coordinates by waving the wand through the capture scene then copying images and waiting for an offline process to extract sphere coordinates followed by the process to generate calibration data resulted in a time consuming process. This wasn't in keeping so a new approach was devised.

3.1.2 New Calibration Implementation

The previous implementation was evaluated for its efficiency, ease of use and the result it produced and a decision was made to retain the underlying method but redo the implementation.

It is desirable to acquire as many frames (and hence sphere coordinates) as the capture process will allow giving the subsequent sparse bundling process the maximum amount of data possible. An exposure of 1200 (previously used) results in 22% less frames being captured and saved in a period of one minute than if 620 is used (see Table 3-1)

Table 3-1 Comparison of Exposure Value to Total Images Acquired

Exposure value	Average FPS	Total images acquired
1200	10	305
620	13	391

3.1.2.1 Wand Development

With the need to detect the balls with high accuracy a new wand with illuminated spheres was developed.

The choice of colour for the balls is important as the cameras sensors are more susceptible to certain wavelengths of light. The Basler cameras present in the Octave research facility (model: piA1000-48gc) have the following colour curves:

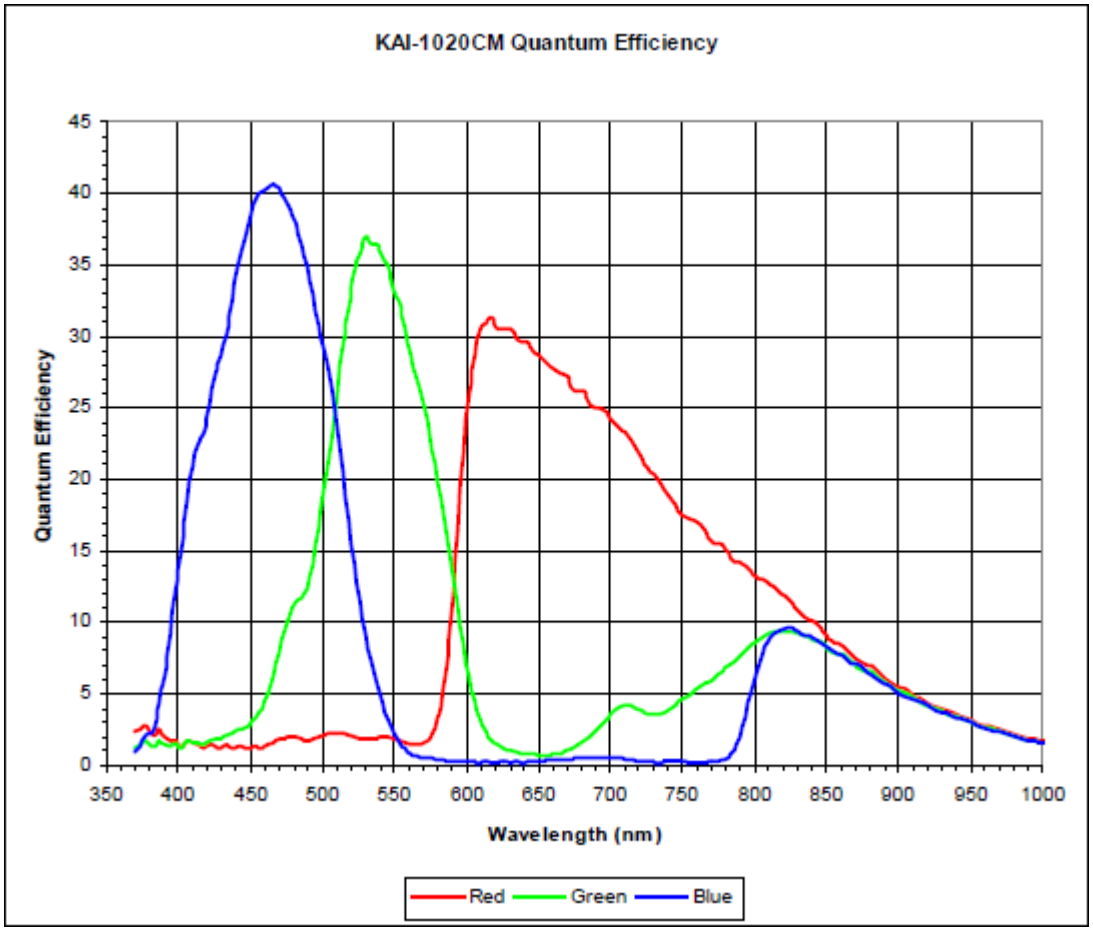


Figure 3-6 Colour Curves for the Basler piA1000-48gc Cameras

The colour curves highlight the sensors heightened sensitivity to red, green and blue. So attempts were made to source illuminating spheres with these colours.

During the initial cycles of iterative development, experimentation using various lamps and L.E.Ds (Figure 3-7), was conducted. Alone, they were not powerful enough to be detected reliably throughout the capture volume.

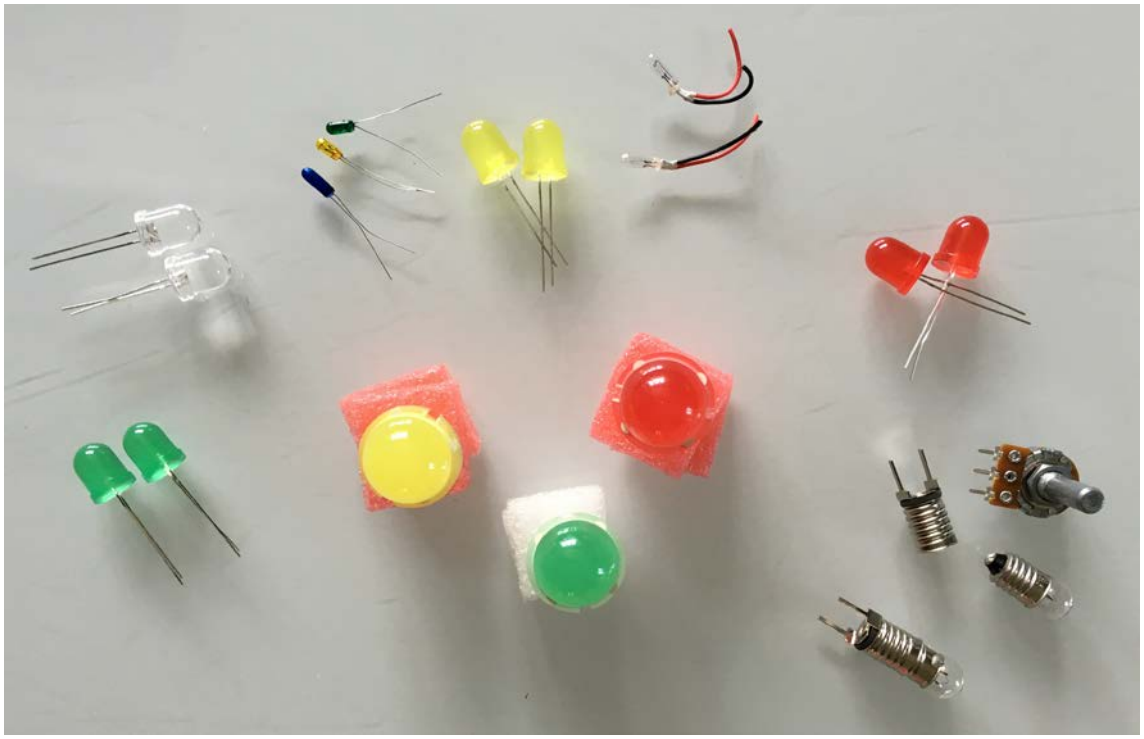


Figure 3-7 Lamps and L.E.Ds used in Early Wand Development Experiments

In an attempt to make the light sources more perceivable they were placed within spheres of various size and colour (Figure 3-8).

Figure 3-8 Spheres of Various Size and Colour used in Early Wand Development Experiments

The spheres were attached to short plastic mounting rods (Figure 3-9) enabling them to be inserted into the wand.

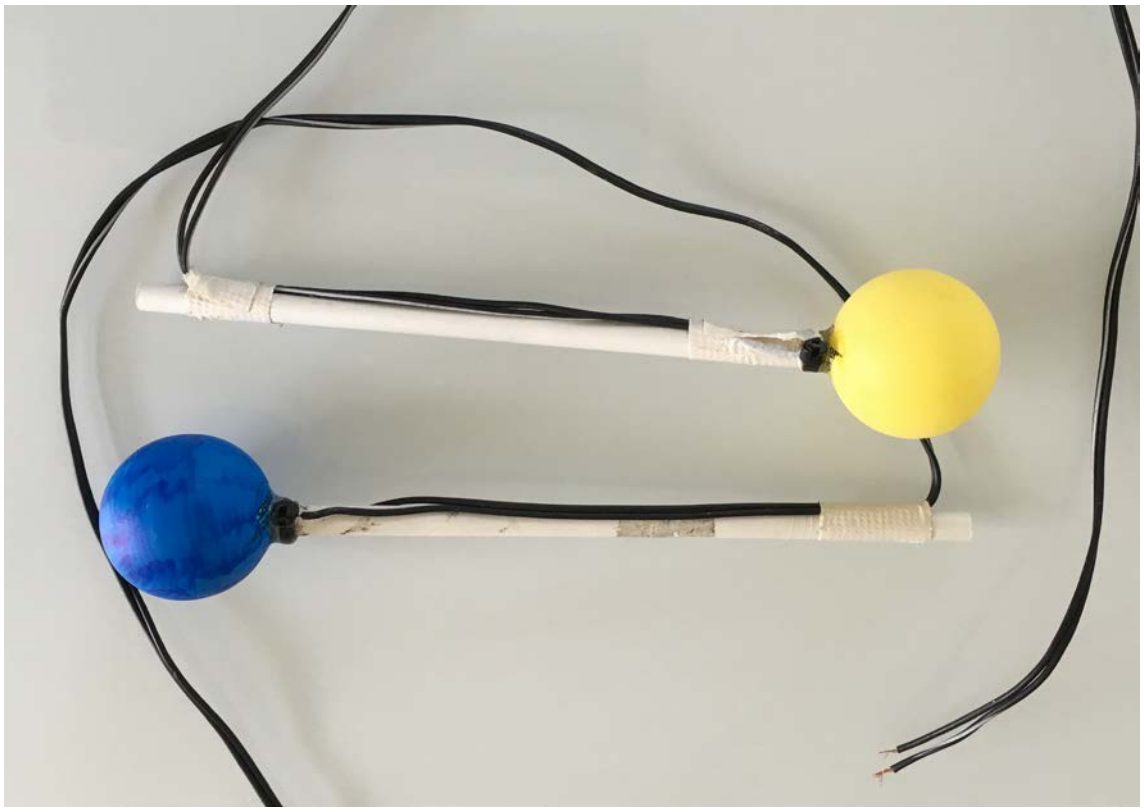


Figure 3-9 Spheres Attached to Mounting Rods

Unfortunately, the adapted spheres were still not perceivable at all locations within the volume either, therefore a new approach to illuminating the spheres was required. This came in the form of pre-made L.E.D illuminated juggling balls. As a blue sphere was not practicably obtainable red and green balls were chosen for the new wand. A grommet was designed and fabricated on a 3D printer. This enabled it to marry with the sphere and attach to an aluminium rod. This rod is then mounted to the wand and adjusted to the desired separation of the spheres. The wand went through several iterations of design, development and testing (as per the iterative deployment model) before the current wand shown in Figure 3-10 was devised.

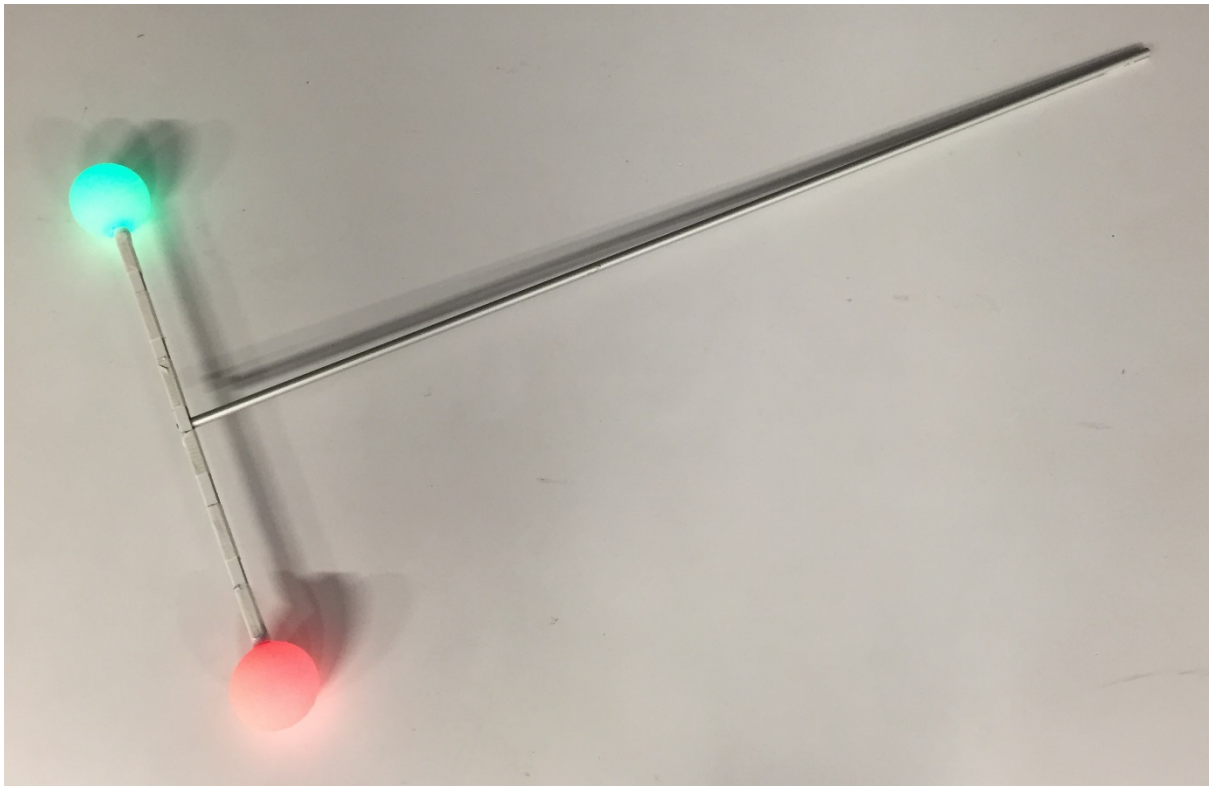


Figure 3-10 Illuminated Sphere Wand

The use of illuminated spheres produces frames containing spheres that are distinct from the background and can be thresholded effectively.

To conduct the calibration the surrounding lighting has to be low but not pitch black and the user performing the waving of the wand no longer has to wear a white overall suit but it is recommended plain clothing without the presence of red or green Figure 3-11. Though it hasn't been tested the calibrators clothing probably wouldn't be an issue due to the low scene illumination.

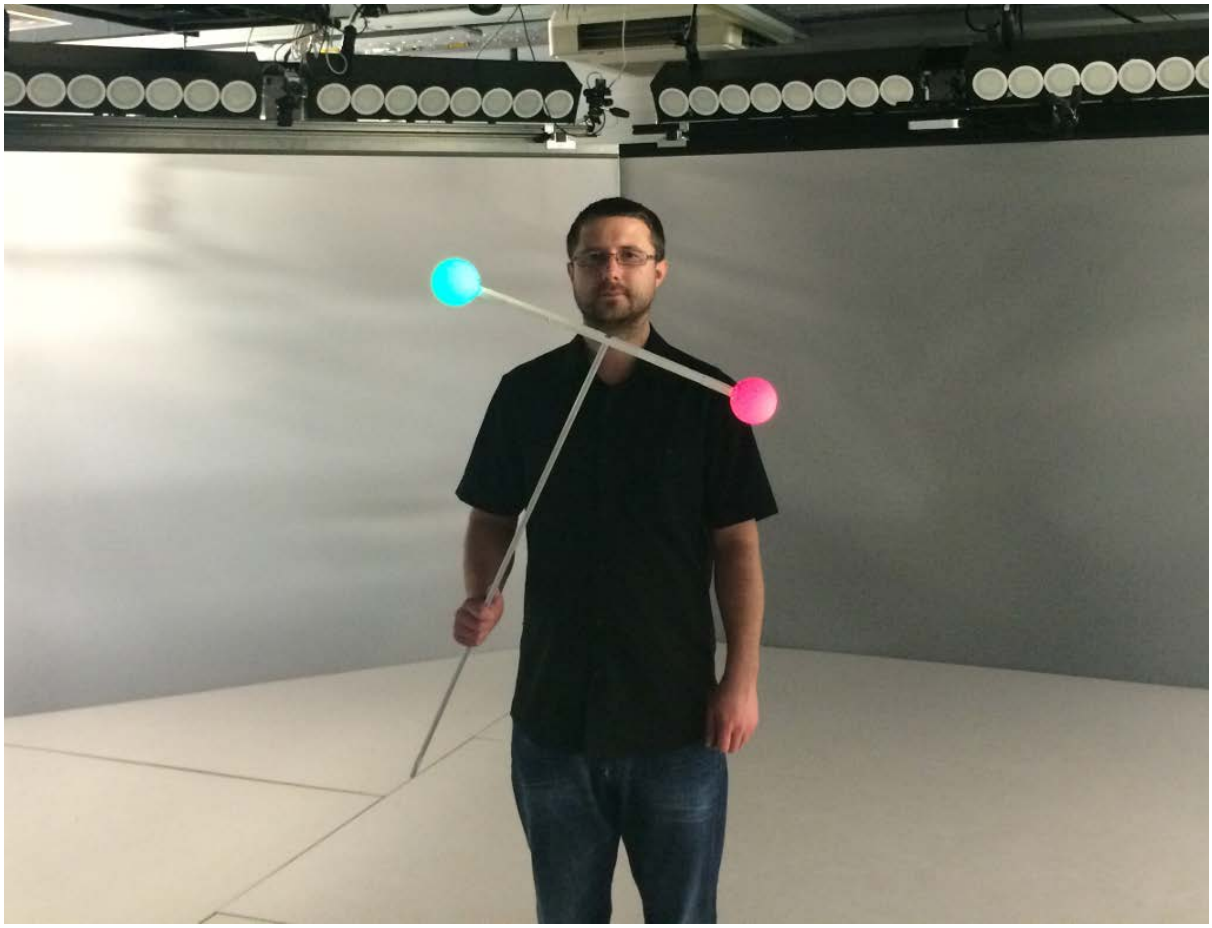


Figure 3-11 Calibrator Holding the Illuminated Sphere Wand

As well as the new wand a new sphere coordinate extractor method was developed to effectively locate the centroid of the balls including when obscured by the wand. As opposed to simply thresholding the image, enclosing the binary blob with a bounding box and using its centre the new method employs several layers of logic and uses the moments of the connected component.

3.1.2.2 Sphere Coordinate Extraction Process

The first stage is to threshold the frame for the hue, saturation and value of a sphere. The colour make-up is predictable and does not require user input to mark the sphere.

Next, the connected components in the thresholded image are detected. If only one component is detected the centre of the moments is used. If two components are detected,

then it is presumed that the wand handle is in the path of the sphere and the camera and circle is fitted to enclose the two components. The centre of the circle is then used as the centre of the sphere. If more than two components are detected, then the sphere is marked as not found. In both cases if the detected component(s) intersect with the bounds of the frame they are also marked as not found. The reason for doing this was the result of testing when it was found that components detected here are not the correct form as they are partially out of view and were affecting the accuracy of the calibration result.

Figure 3-12 below highlights results from the new method:

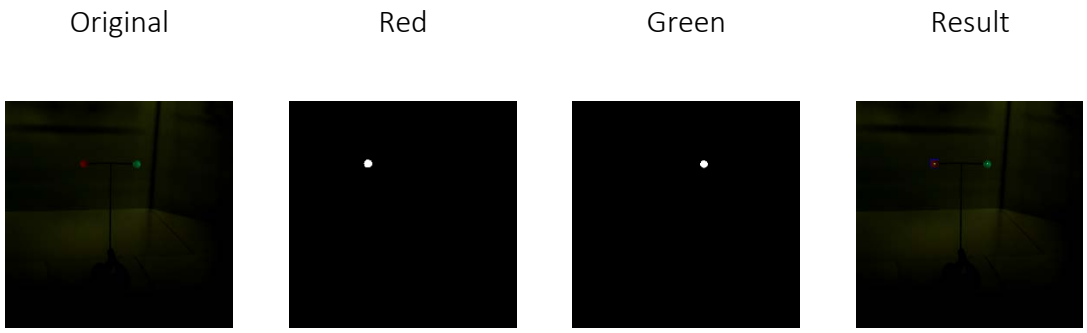


Figure 3-12 Result of Sphere Detection with the New Method

In Table 3-2 the increase in images that can be acquired by reducing the exposure is shown.

Yet saving images to disk is a time consuming process and reduces frame rate. Furthermore, copying the images and then processing to detect the spheres offline adds additional complexity and delay. With a goal of a useable system in mind several updates were made to simplify the calibration process for the user.

First, the sphere coordinate extractor was integrated with the capture software. This facilitated numerous things including a marked improvement of the frame rate possible:

Table 3-2 Comparison of Exposure Value to Total Images Acquired

Exposure value	**Average FPS**	**Total frames processed**
620	35	1131

This results in a ~73% improvement in the amount frames than in the previous implementation.

Additionally, the capture software was updated to enable users without specific knowledge to use it for calibrating the system and new mode of operation was added to the capture software that starts it with a configuration specific to calibrating. The mode can be toggled in an XML configuration file on the lead capture node with all other nodes being informed of the mode during their initialisation.

The capture nodes record the coordinates of the spheres to a central location so there is no requirement to copy files from each node.

Post-acquisition of the sphere coordinates the multi-camera wand calibration software is executed. The software has been updated to output the exact format required by the 3D reconstruction process thus reducing the requirement of another challenging manual step in the previous implementation.

The flow diagram for new procedure is shown in the following diagram (Figure 3-13):

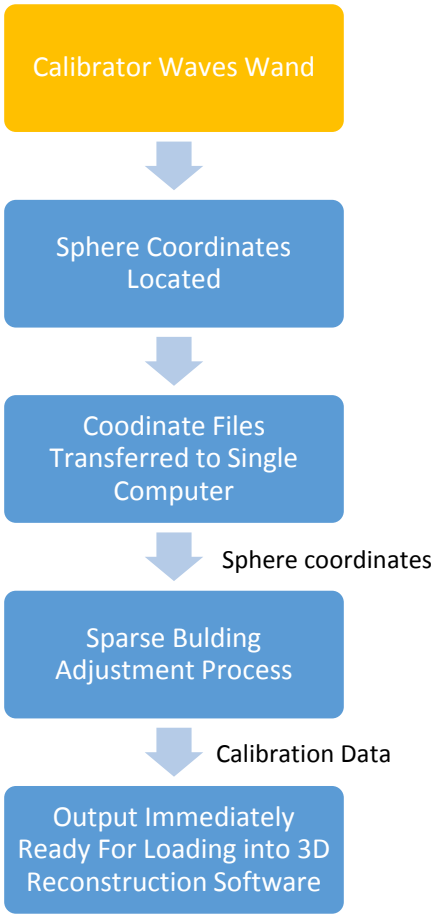


Figure 3-13 New Calibration Process (Blue: Automated, Orange: Manual Process)

For completeness documentation is provided to communicate the calibration process and is included in Appendix A.

The centre of the spheres was originally determined by locating the centre of a bounding box. Then, in a later revision by finding the centre of all moments. Further evaluation of results during the iterative development highlighted issues with the shape of spheres on the edge and these were removed.

The following table presents actual results from the different methods apply to sample sets of 1500 frames:

Table 3-3 Results from Different Sphere Detection Methods (Lower Values Better)

Method	Reprojection Error
B. Box	1.14168
B. Box (outlier removal)	0.732014
Moments	1.0702
Moments (outlier removal)	0.702618

3.1.2.3 Summary

The 2D pixel re-projection error defined as the RMS error is an excellent indicator of the level of success for a calibration. The pixel re-projection error can be used to determine the calibration quality of an individual camera. The new implementation results in a consistently low RMS error rate and the quality of the calibration result is to a consistently high standard. The time taken to achieve a successful calibration is markedly reduced when compared to the previous implementation. The results of the development enables researchers to experiment with updated camera poses and recalibrate in a simple and timely manner.

3.2 Colour

The visual quality of the 3D model is greatly affected by the colour consistency across the multiple cameras used in its acquisition and faithful reconstruction affects believability. This chapter outlines the issues and presents a solution to the problem.

Incorrect colour calibration across multiple cameras was highlighted as an issue with the previous work. Not only did some cameras produce an output that was markedly different, they were all over exposed (and by different amounts) as a direct result of the requirements for successful segmentation (detailed in the following chapter). This is not just an issue limited to the cameras in the Octave Research Facility, in fact, it could be an issue with any type of camera. Furthermore, the previous approach to multi-view rendering used the images with no blending (Roberts et al., 2013) and while this resulted in clear eye and face representation, it had an undesired effect of producing visible lines in the border of different images.

Believability of the reconstructed model is dependent on the quality of the texturing process and it transpires that manual selective texturing has been used in the past to produce results. The work done and detailed in the following sections enables all textures to be used improving the quality of the model especially when running live. The striping effect with the previous rendering implementation (Figure 3-14) and a noticeable change when moving viewpoint across cameras in the new render client is notably reduced.

Figure 3-14 Striping Effect

Initial tests on the inherited framework revealed a problem within the system that manifested itself by producing inconsistent output, i.e. mismatched textures on the final 3D model. A brief analysis of a selection of sample texture sets captured by each of the cameras at precisely the same time lead to the following conclusions:

a) Each camera produced varied results in terms of luminance and saturation when compared to other cameras.
b) The results produced by each camera appear to be stable across the set and did not shift; any texture produced using a given camera appeared to be consistent with all other textures produced using this camera.
c) The inconsistency was produced either by the camera hardware of its proprietary control software.
d) Ambient light could be a contributing factor to the problem.

A decision was made to reset and modify all camera settings using their proprietary control software and conduct a series of tests. Settings modifications would be global (each change would involve offsetting all respective camera settings by exactly the same value).

The tests involved placing a professional colour calibration card in a central spot of the Octave installation (Figure 3-15) and capturing sample sets of textures from all cameras under three different camera settings, i.e.

a) Factory settings; this step was necessary to establish a base result.
b) Low frame-rate settings consistent with ambient light frequency; this step was necessary to rule out the possibility that the ambient light flicker frequency (50 Hz) may have been interfering with the frame rate. The capture frame rate was set to 25 FPS globally, or half of the supposed light flicker rate, to cancel out any interference.
c) High frame-rate; the aim of this step was to amplify the potential interference between the potential ambient light flicker frequency. The frame rate was set to the maximum possible value of 48 FPS.

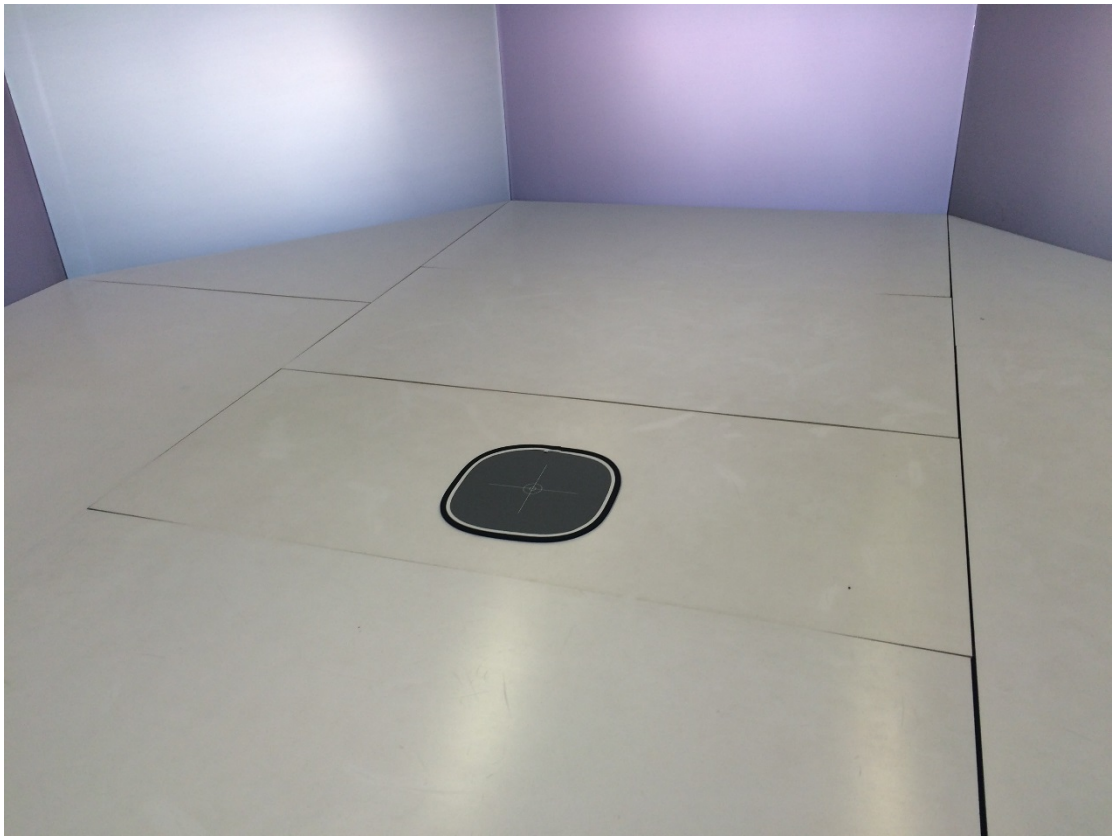


Figure 3-15 - Professional Colour Card Positioned in Centre of Octave

The resulting sets of textures were compared using Adobe Photoshop. The test yielded the following conclusions:

a) Resetting the cameras to their factory settings using their proprietary control software improved the output consistency greatly, both in terms of colour as well as luminosity.
b) Ambient light flicker rate did not produce any detrimental effect in terms of consistency, as no evidence of interference was found.

The above tests were also conducted using other objects of varied size and complexity. Each test yielded the same results.

It was concluded that the initial output inconsistency in the inherited framework was mainly the result of mismatched camera settings on the proprietary control software level. These mismatched settings were the direct result of attempts to attain successful

segmentation with previous background segmentation method detail in Chapter 4.1.4. Some inconsistency still remained after the factory reset, however the overall output results were vastly improved. However, whilst conducting the testing above it was determined that two of the cameras did have a markedly different output. The difference was not colour but luminosity. It was extremely high compared with the other cameras as can be seen in Figure 3-16 below.

Figure 3-16 - Luminosity Inconsistencies Between Two Cameras

Although unlikely it was possible that light source positioned overhead in close proximity to one of the cameras could be causing the issue. To rule this out the two cameras with the most different output were swapped and the fault remained with the camera in its new position. This was determined to be a hardware fault out of our control.

To correct for this fault and to add further value to the system (as web cams and Kinect camera support has now been added) some post processing was integrated into the pipeline to enable the fine-tuning of brightness post acquisition. Originally a two-step process was proposed and tested:

3.2.1 Advanced Colour Correction Method

First a brightness correction was performed to compensate for inconsistencies that could not be resolved by manipulating cameras hardware settings.

Secondly a hue correction was performed to correct any remaining inconsistency issues resulting from the fact that the colour output was marginally different with each camera.

3.2.1.1 Brightness Correction Process

All images loaded on one screen using a calibrated monitor (Figure 3-17).

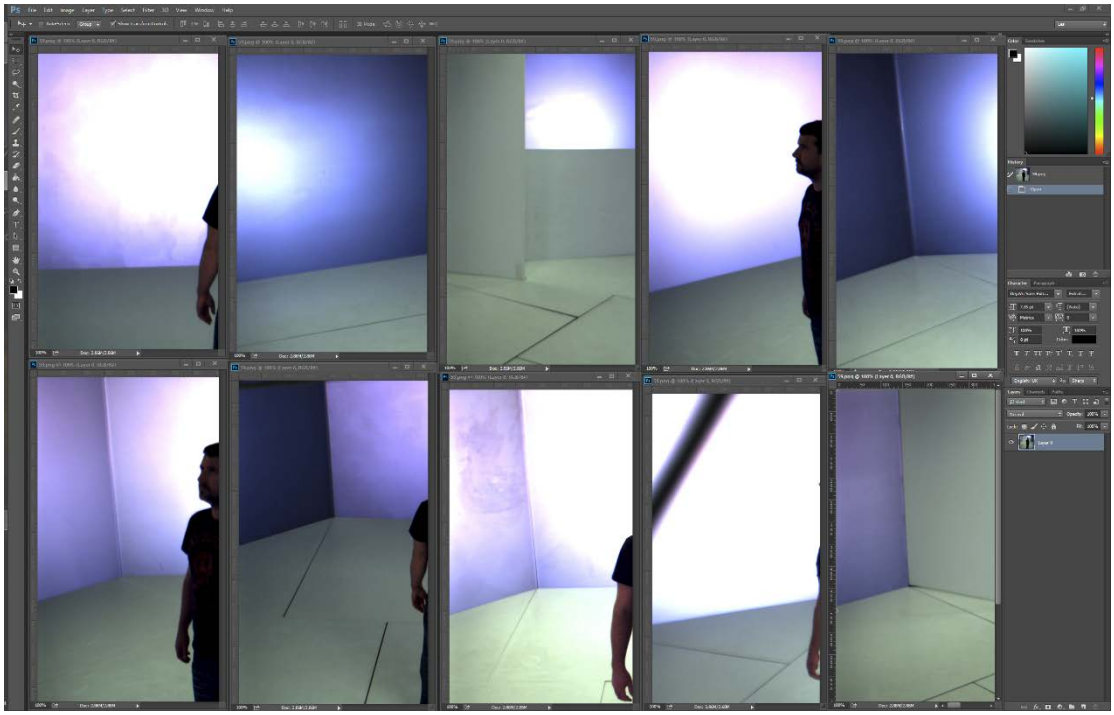


Figure 3-17 - Syncronised Images from 10 Cameras

A point of reference across all images was chosen based on the following criteria:

- colour uniformity: it was crucial that the object's colour be the relatively uniform across its surface as registered by each camera to allow for accurate comparison
- the amount of light reflected from the object: the object had to be bright enough for its colour to be clearly visible

- position: the object had to be clearly visible from each camera's perspective

It was decided that the human head met the above criteria. The images were zoomed in on the face as shown in Figure 3-18.

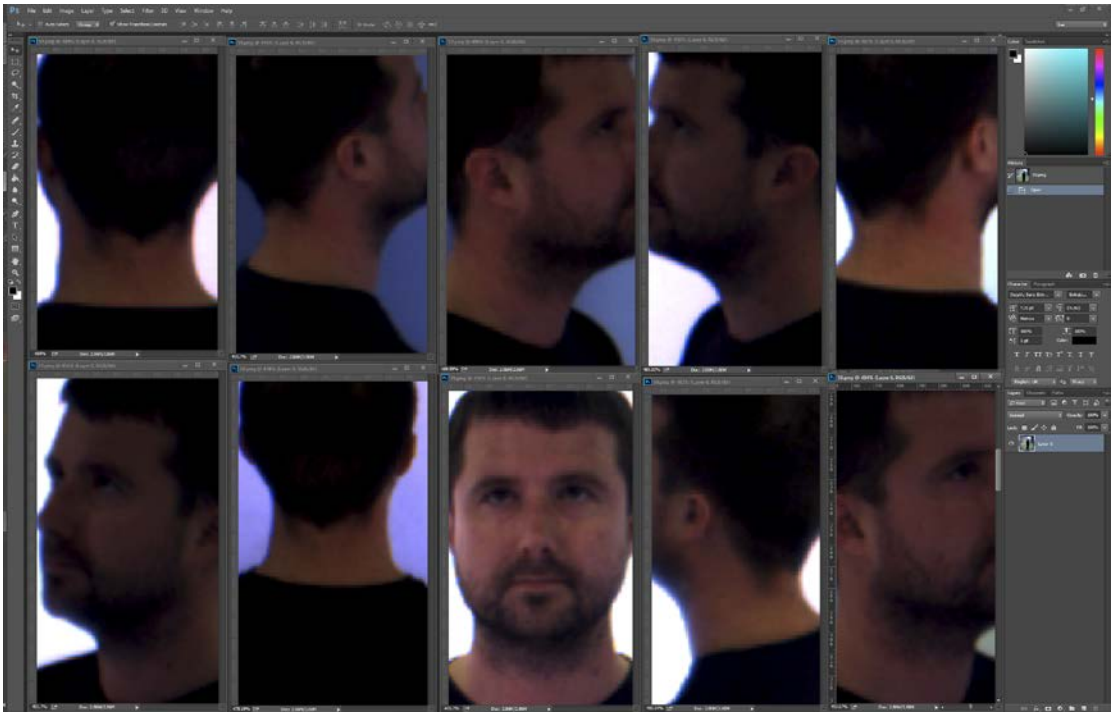


Figure 3-18 - Syncronised Images Focused on Head of Participant

Each image was adjusted individually for brightness (Figure 3-19).

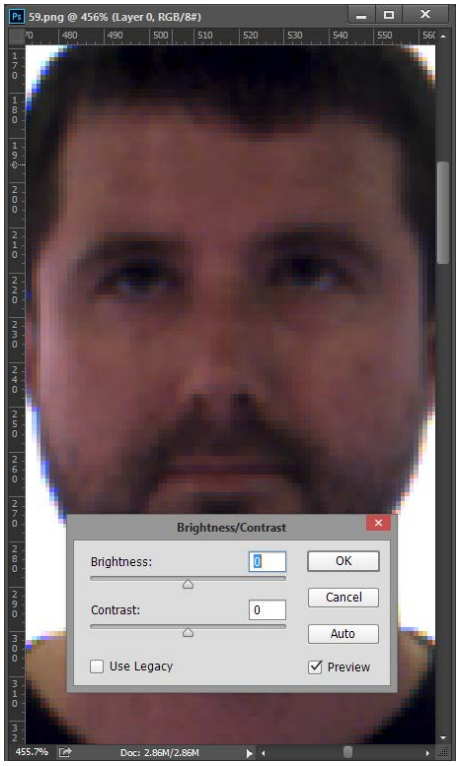


Figure 3-19 - Brightness Adjust Taking Place

The final result is shown in Figure 3-20 below:

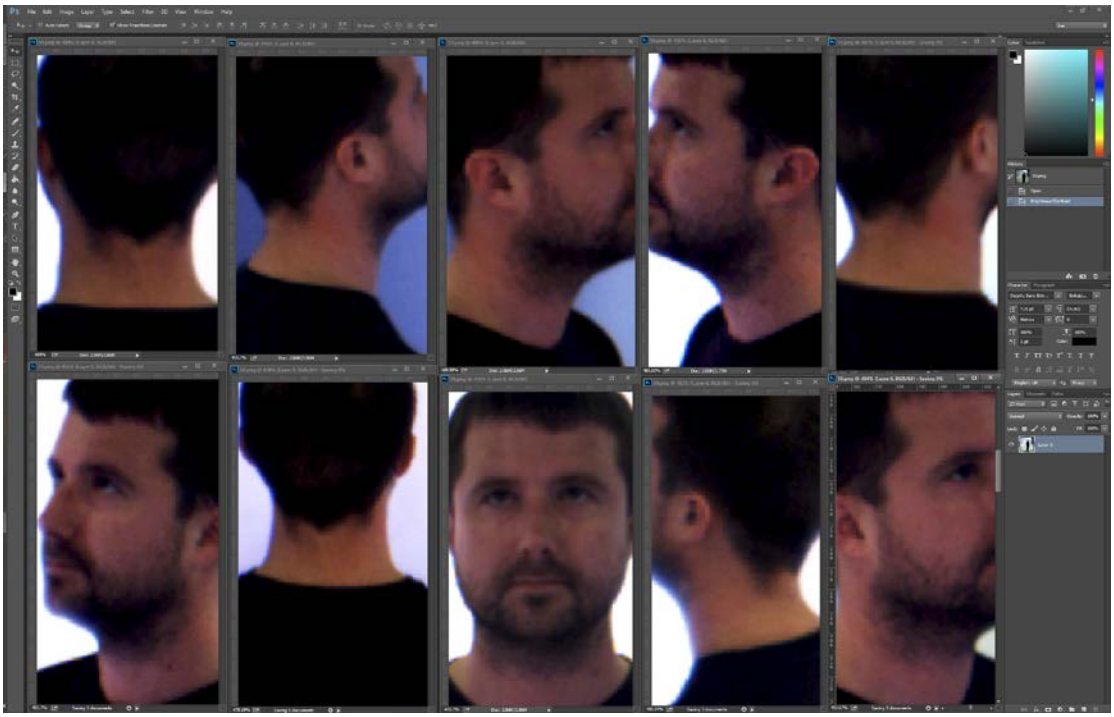


Figure 3-20 - Result of Brightness Correction

3.2.1.2 Hue Correction Process

The purpose of this post-processing stage was to equalize the overall colour make-up of all respective output images produced by all cameras. In order to do so, a five-stage method of image analysis, comparison and adjustment was devised.

Stage 1

The first stage involved simplifying the image via the process of posterization, in order to extract three most prominent colours from each respective image.

Stage 2

The second stage was to record the value of each of the resulting colours and assign each to one of the following three categories based on their general brightness: “low” (for the darkest colour), “high” (for the brightest colour) and “medium” (for the remaining third colour). For more accuracy, the level of posterization could be set to produce more colours. The brightest and darkest colours would still be filed under “high” and “low” categories respectively, while all other colours would be assigned to intermediate “medium x” categories, where “x” would be a decimal number marker assigned to a given colour based on its overall brightness, e.g. “medium 2” colour would be brighter than “medium 1” but darker than “medium 3”.

Stage 3

The third stage entailed comparing the resulting colours from one user-selected camera against the corresponding output of all other cameras. The camera selected by the user would be known as “master camera”, and the set of colours extracted from image output of that camera would be known as “master set”.

The “low” and “high” colours may often be captured as pure black and pure white respectively due to under or overexposure issues beyond viable control, regardless of the

overall colour make-up; an underexposed part of the image can often be recorded as pure black, whereas an overexposed part can be recorded as pure white. Comparing colours sampled from these parts of the image would therefore not be accurate. For this reason, the “low” and “high” categories were to be excluded from the colour comparison process.

Stage 4

In the fourth stage the corresponding “medium colours” from all cameras were changed to match those recorded by the “master camera”. The modification was to be global for each set of colours extracted from a given image, i.e. change applied to “medium 1” colour would also affect all other “medium” colours within the image's extracted colour set.

The change necessary to adjust a given set of colours to match the “master set” was recorded.

Stage 5

In the fifth and final stage the recorded colour change was applied to the original corresponding images (Figure 3-21).

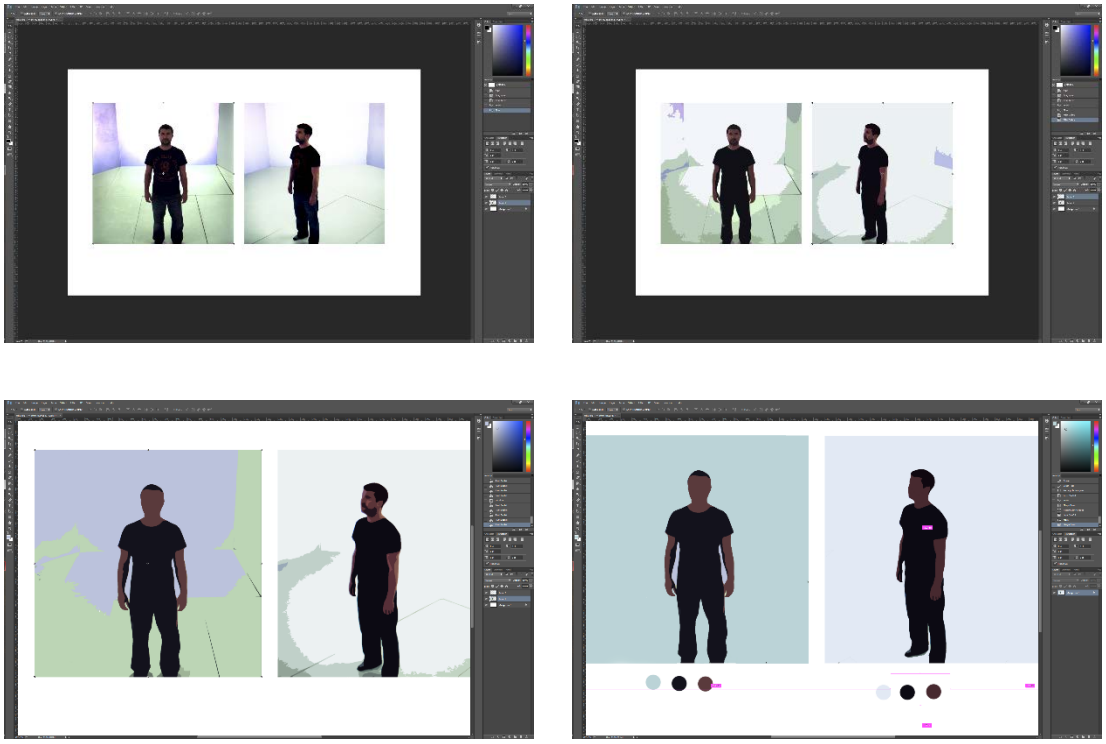


Figure 3-21 - Hue Correction Process

Although the result from using both steps of the process is best the second stage current requires human input. It could be possible to automate but given time constraints and the likelihood that the process performed on currently available hardware would hinder the real time performance it isn't integrated into the complete system.

3.2.2 Simple Colour Correction Method

Brightness and contrast correction has been fully implemented and integrated with the system. Although this does actually alter the hue it is still named colour correction to avoid confusion. It requires the user to observe the professional colour calibration chart from each camera and refine the values as necessary.

In addition, a new camera setting saving and loading feature removes the laborious task of setting camera settings manually and enables different profiles to be created and loaded with ease

3.2.3 Example Colour Correction Results

Example results of the colour correction processes are shown in Figure 3-22 below. The image on the left is the uncorrected 3D model, the centre image shows the result of the simple correction and the right image the advanced correction process.

Figure 3-22 - Colour Correction Results

Here are some resultant images taken post capture from the recorded live 3D video session (Figure 3-22and Figure 3-24):

Figure 3-23 3D Avatar of User Stance A

Figure 3-24 3D Avatar of User Stance B

3.2.4 Summary

This chapter highlighted the effect colour inconsistencies across cameras can have on the faithfulness of the reconstructed form. It then demonstrated a systematic approach to determine the issue facing the camera setup in the Octave Research Facility. Then it presented two methods to correct for colour inconsistency across cameras. The first approach was advanced and required some user input and while it was not integrated into

the processing pipeline is of merit to describe due to the quality of the results that can be achieved. The second method improves the result and has been fully integrated into the processing pipeline in such a way that it can use a single configuration file conveniently stored on a single capture node which then pushes the correct settings to each of other captures nodes.

Chapter 4

Background-Foreground Segmentation

This chapter presents the background-foreground segmentation improvements that have been contributed to the system. It begins by analysing the current methods employed in the prototype to gain an understanding of the challenges and current shortcomings. Then it presents the steps taken to improve results in visible light and finally demonstrates an approach to perform segmentation in Infrared light.

4.1 Visible Light

The reconstruction process uses shape-from-silhouette to form the visual hull for the 3D model therefore the quality of the model is dependant in part on the quality of the silhouettes. A process known as background segmentation can be used to create the silhouettes. There are numerous approaches to performing background segmentation the simplest being taking a reference frame and then looking for differences in later frames classing them as foreground pixels. Inevitably this simple approach suffers from noise and changing conditions in the environment. Many solutions have been devised to deal with these issues. The segmentation algorithms used in the previous work were:

- GPU-Based Foreground-Background Segmentation using an Extended Colinearity Criterion (Griesser et al., 2005) which was based on the work of (Mester et al., 2001) and will be referred to as the 'Mester Implementation'
- Chroma-keying (van den Bergh & Lalioti, 1999)

In practice in the conditions in the Octave they have shown various shortcomings:

4.1.1 Mester Implementation

It was difficult to achieve optimal conditions required for accurate segmentation with each camera requiring a different set of parameters. The different parameters had a direct effect on the colour consistency of the generated model. Noise, shadows and floor movement of the Octave environment were often incorrectly segmented as demonstrated in Figure 4-1 and Figure 4-2 below:

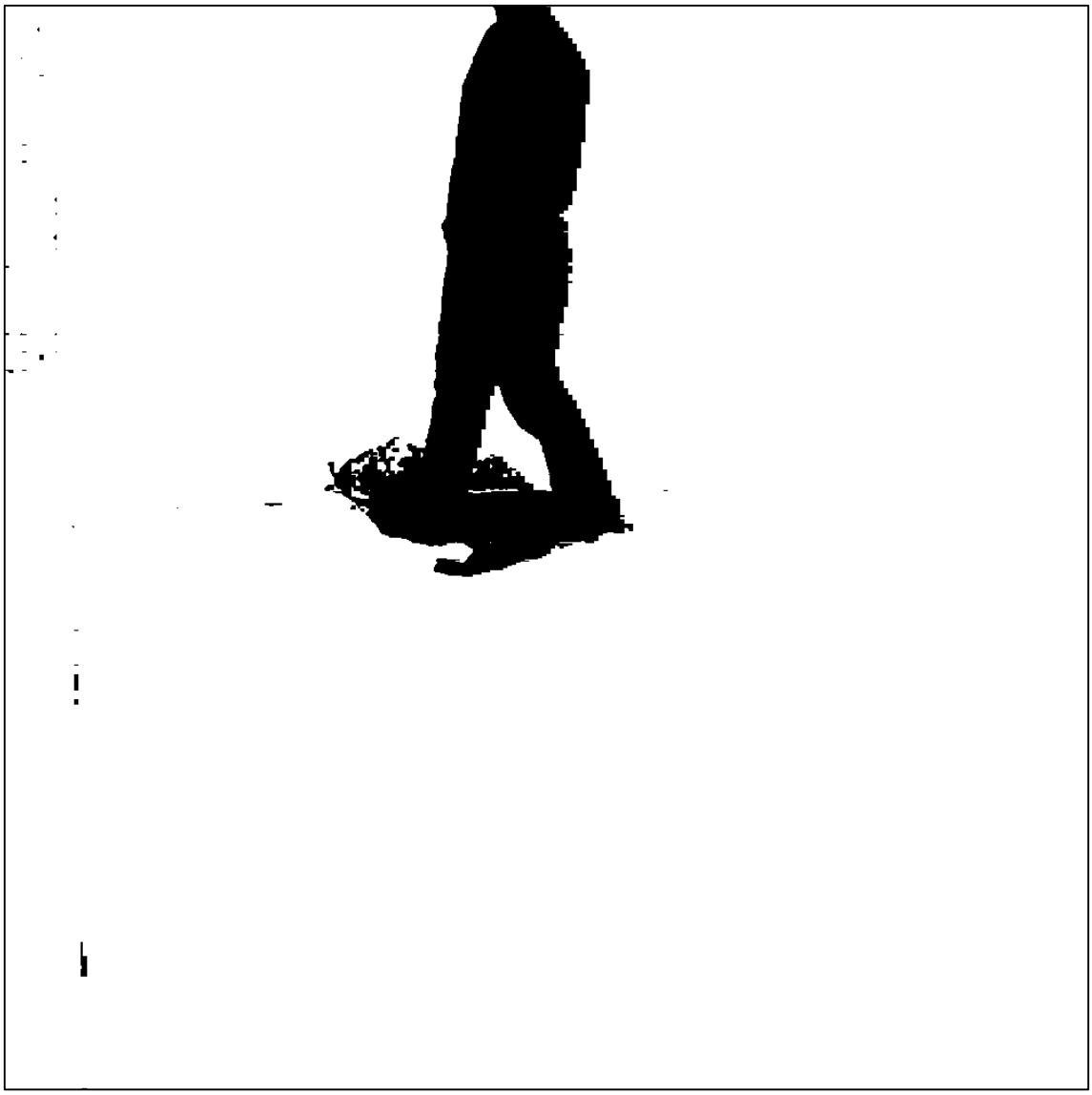


Figure 4-1 Poor Segmentation Result with Misclassified Shadow

Figure 4-2 Poor Segmentation Result with Noise

To achieve a reasonable level of quality segmentation the cameras required different colour settings, longer exposures and high gain (to make the background as uniform as possible) resulting in unnatural colours on the reconstructed 3D avatar.

4.1.2 Chroma-Keying Implementation

Using the Chroma-keying implementation resulted in participants being bathed in the light colour chosen to segment against making them appear unnatural. An example of this

effect is shown in Figure 4-3 where the participant appears with a green tinge thus demonstrating that for the best 3D model users should be illuminated with white light.

Figure 4-3 - 3D Model Generated Using Green Chroma-keying Method

4.1.3 Summary of Previous Segmentation Implementations

The methods also employed erosion which results in loss of image data and hence information. Also the many settings in Mester for each camera - that must be configured via a trial an error process - make the system inhibit ease of use especially when moving the cameras requires the process be repeated.

To compensate for noise, the previous implementations of background segmentation employed a global erode function on the silhouette image directly after the segmentation

process. Whilst this is generally successful at removing noise it has a direct consequence of losing image information in the process as it does not discriminate and erodes the silhouette of the user being captured too.

The problem can be quantified by measuring:

- The deviation from the ground truth silhouettes for each of the different methods
- Observing the resultant 3D model quality
- Analysing temporal quality as poorly segmented participants result slower hull reconstruction

4.1.4 Updated Segmentation Implementation

The shortcomings of the previous approaches to segmentation resulted in them being determined as not fit for purpose didn't coincided with the aims of this system.

As the previous two approaches only worked with either unnatural colour on the participant with Chroma-keying or exposure and high gain the decision was taken to find a method of segmentation that worked with the cameras configured in a manner that would attain the best colour properties for the resulting 3D model i.e. as natural as possible. It should also be robust, accurate and require as little user intervention during setup as possible, preferably none. The process of achieving this was as follows:

To begin with all cameras are configured to have as faithful and consistent a colour across the array as possible utilising the colour correction process implemented in Chapter 3.2.2 if necessary.

Following this sample datasets of people walking through the capture volume and performing gestures such are waving are acquired.

The sample datasets are processed with various background-foreground segmentation algorithms using a tool named BGSLibrary (Sobral, 2013) and the results analysed.

An exhaustive analysis of each method is not necessary is not provided. Rather a summary of findings.

This analysis highlighted several issues with algorithm selection:

- Background learning rate
- Sensitivity
- Ambient light level requirements
- Shadow detection

4.1.4.1 Background Learning Rate

The rate at which the model adapts to changes in the video image. Low values correspond to a slowly adapting model. High values make the model adapt quickly to scene changes. Users becoming part of the background.

4.1.4.2 Sensitivity

Determines the sensitivity to changes in the background. Low values enhance the detection of objects in the scene, but also make the model more sensitive to noise.

4.1.4.3 Ambient Light Level Requirements

Refers to the amount of available light present in the scene. It is an important factor to consider when choosing an approach to segmentation. Equally important is the distribution of the available of light across the scene.

4.1.4.4 Shadow Detection

Implementations featuring a shadow detection implementation. It can be built in to the algorithm or executed post segmentation.

4.1.4.5 Gaussian Mixture Model Results

Based on the results of the experimentation a Gaussian Mixture Model based implementation of background subtraction (Zivkovic & van der Heijden, 2006) was elected to be implemented. It will subsequently be referred to as the Method of Gaussian (MOG) implementation. Noise that remains after the MOG segmentation is consistently low compare with the previous Mester implementation. Any remaining noise is removed using a de-speckle process rather than a global erode to ensure that the silhouette isn't modified. Noise induced by the floor seams of the experimental environment moving (shown in Figure 4-4 below) is removed by removing connected components that are much smaller than the largest components present. It should be noted than when the floor seams are taped over this problem doesn't occur but due to the fact that the experimental environment is configurable and used for other research it is not guaranteed that the tape will be in place.

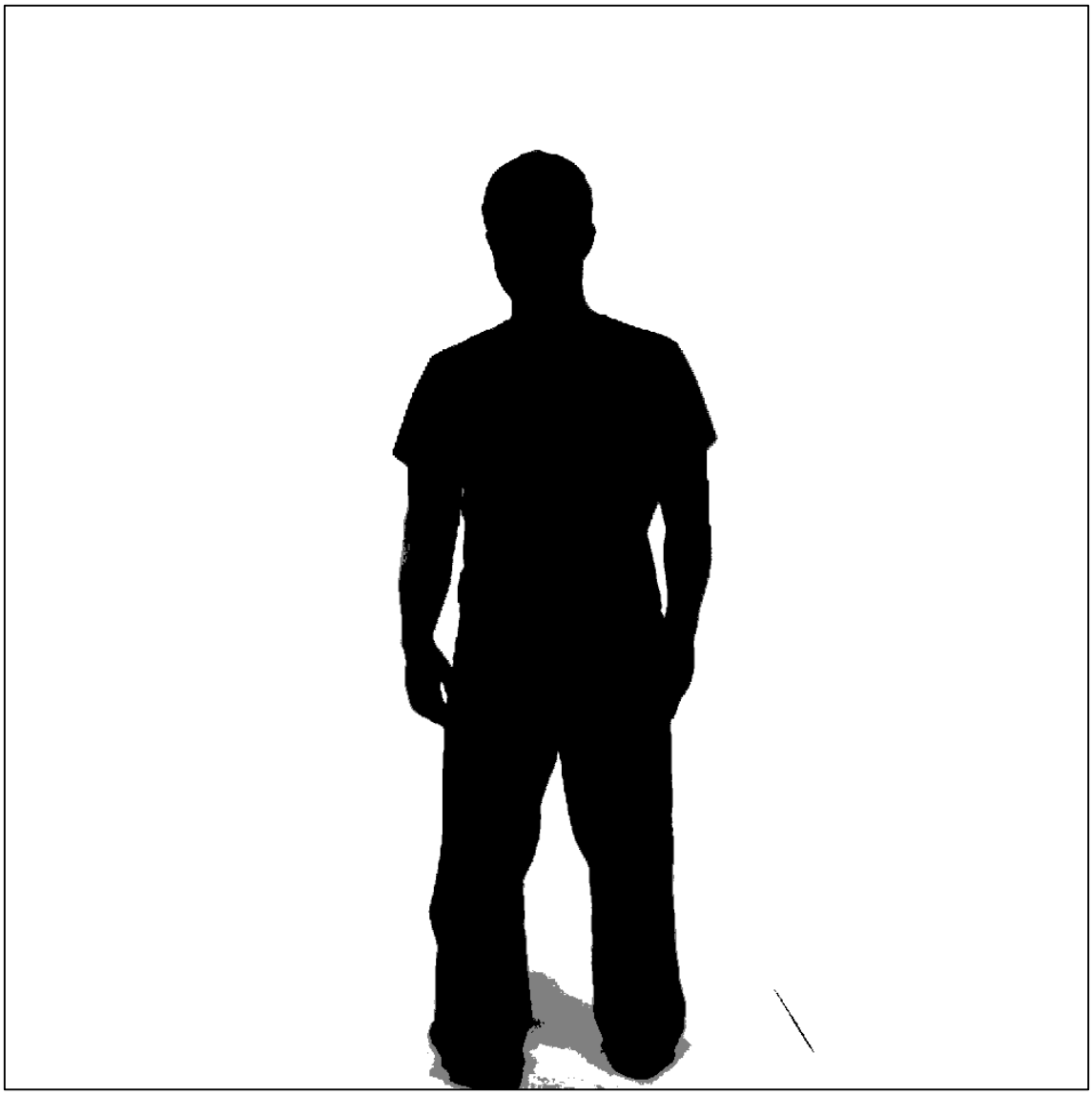


Figure 4-4 - Silhouette Image Demonstrating Segmentation of Seams in Octave Floor and Detected Shadow

Below, in Figure 4-5, is an example result from the MOG method employed, the area detected as shadow is highlighted in grey and will be removed before the silhouette is sent to the 3D reconstruction process.

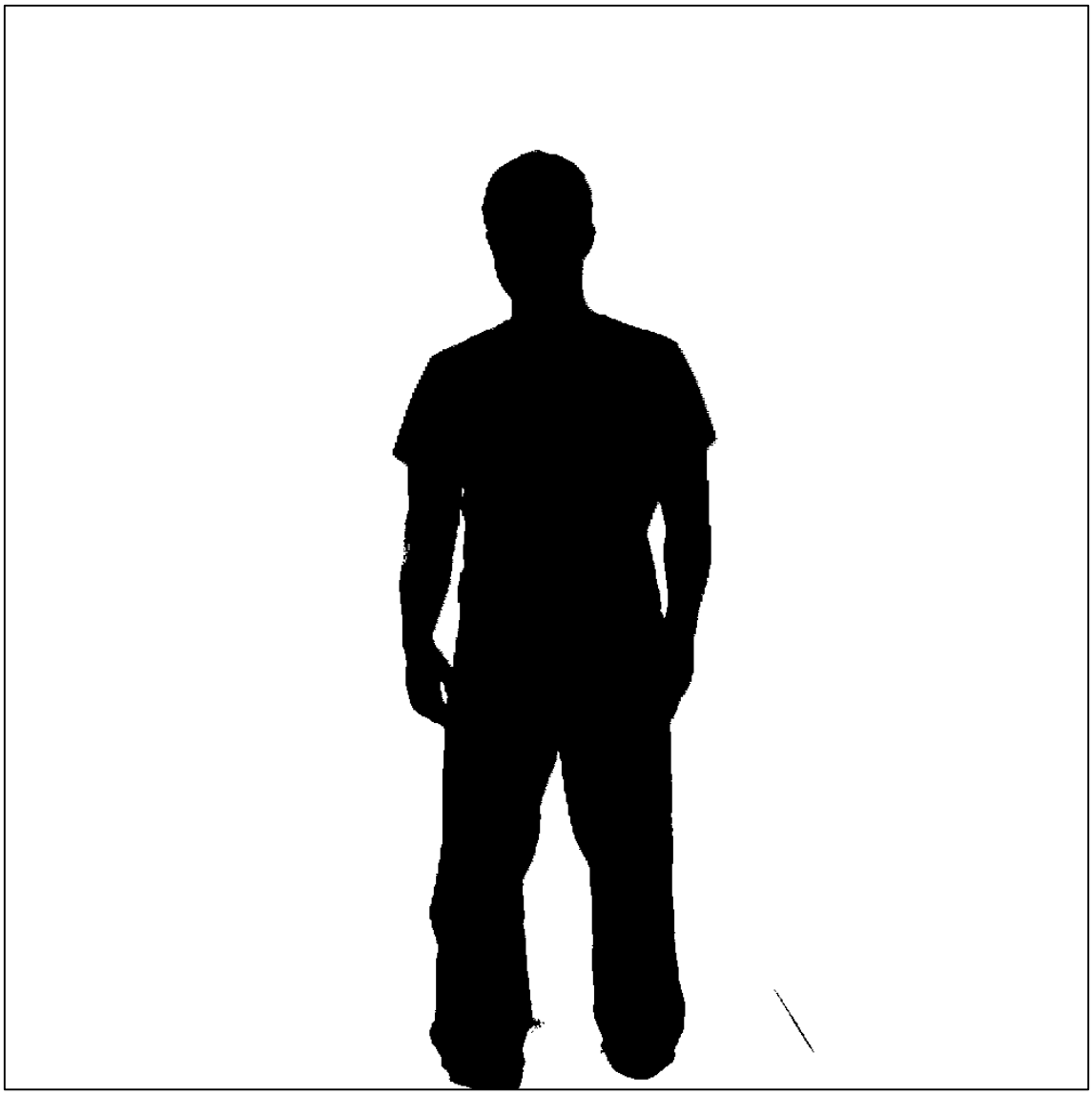


Figure 4-5 - Silhouette Image with Detected Shadow Removed

The final result after the smaller components have been removed is shown below in Figure 4-6:

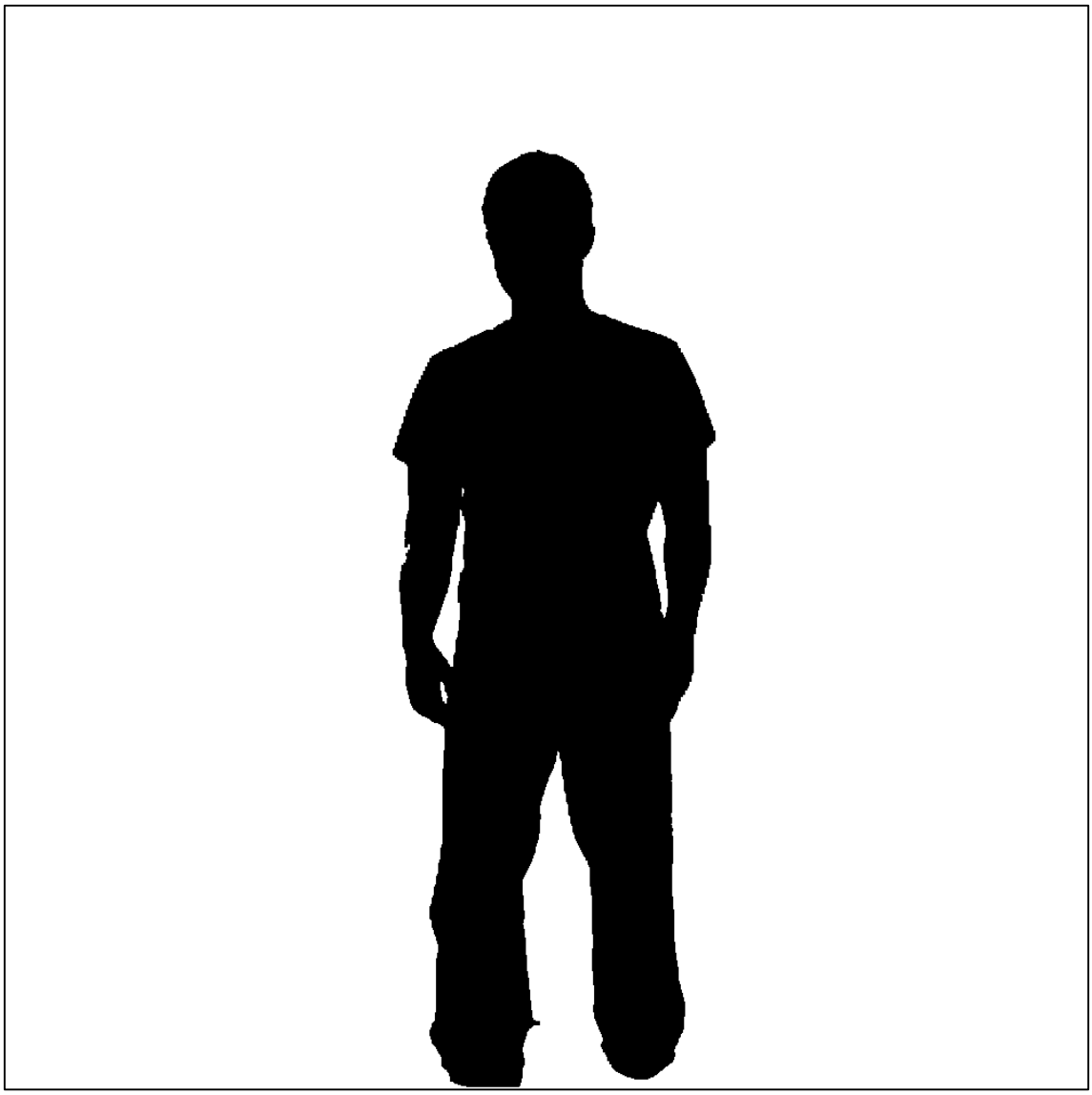


Figure 4-6 Final Silhouette Image with Shadows and Noise Removed

The Figure 4-7 below highlights the improvement of better segmentation on the extremities. On the left is the hand generated with a set of silhouettes produced using the Mester process. It is clearly slightly deformed and web like. The centre image is the result of using the new MOG approach and it is a much more faithful reconstruction. Finally, the right image demonstrates what can happen if an erode process is passed over the already accurate MOG result - parts of the silhouette are reduced too greatly and as is clear one finger is now missing and the others have sections missing.

Figure 4-7 - 3D Hand Result using Mester (left), MOG (centre), MOG with a Single Erode Pass (right)

The improvements result in a more faithful 3D reconstruction. The only downside is that on comparable hardware the MOG takes longer to process each frame than the previous Mester. However, when timings are considered for the complete pipeline of the system not just segmentation the overall result is quicker. For example: in the inherited framework the GPU implementation of the Mester algorithm was used to segment the foreground from background. The performance was around 0.5 ms per frame and when looking solely at the component individually this is a relatively fast process. However considering that the quality of the segmentation has a direct impact on the next process in the pipeline, contour encoding, it is actually quicker overall to spend greater time getting a better segmentation result so that this time is reduced. This is what transpires: a typical contour detection and encoding time with a segmented image from Mester is 10ms. The time to segment with MOG implementation is typically 5ms yet the contour encode is 2ms so the time actually saved is 10.5 – 7 = 3.5ms. This is only the time saved for these two components. The 3D recon time is also slower with poorly segmented images.

4.1.4.6 Visual Improvements

The following figure (Figure 4-8) shows a side-by-side comparison of the two methods of segmentation. It is evident that MOG method captures more information than the Mester and results in a reconstruction that is more faithful.

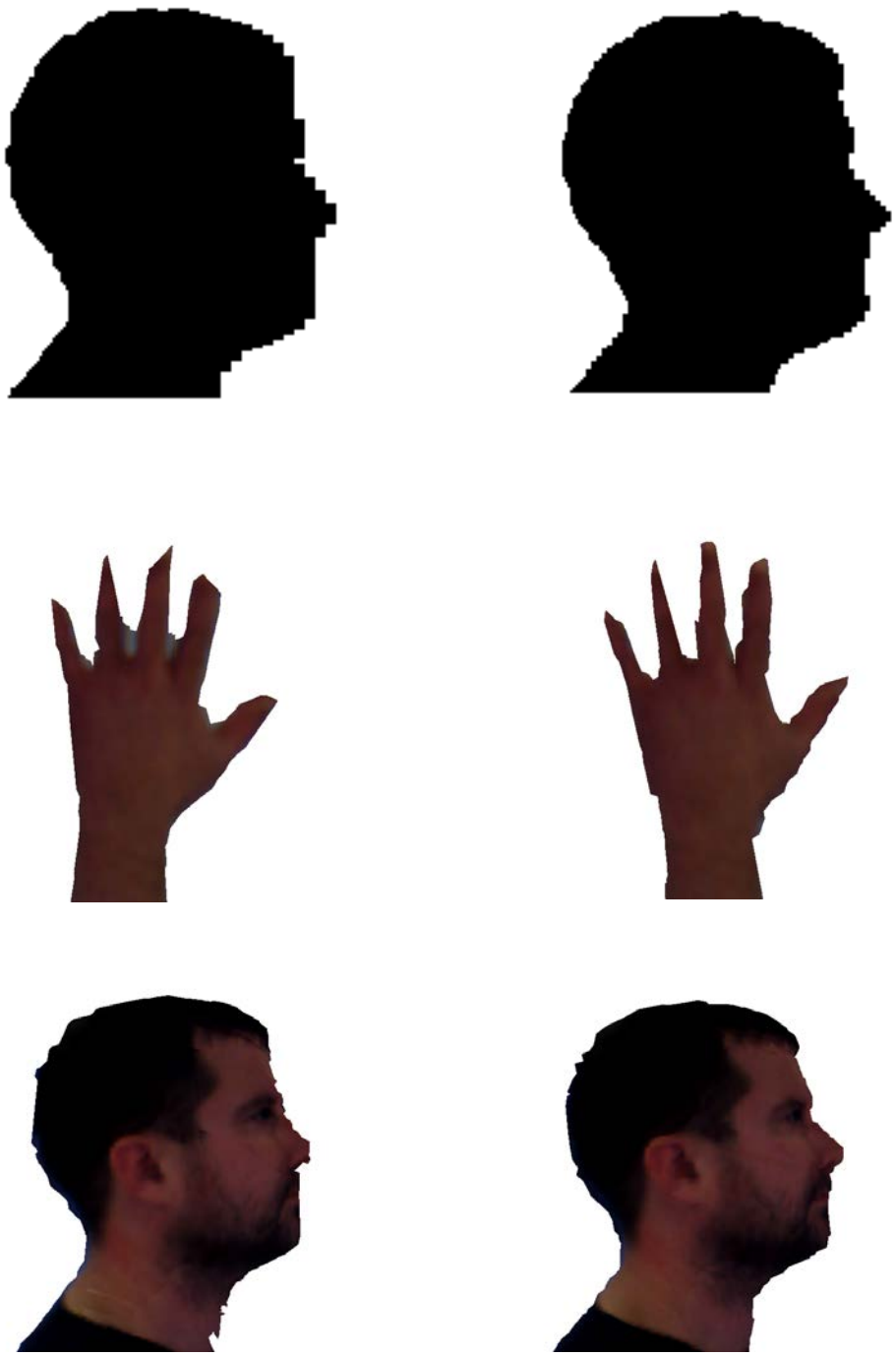


Figure 4-8 Mester (left), MOG (right)

4.1.4.7 Ground Truth Evaluation

To further validate the results the silhouettes from both methods were compared to the Ground Truth silhouette. This process is not just applicable to the methods used in the thesis but could be used with different methods in the future.

To acquire the ground truth silhouette an advanced software application capable of annotating documents named Aletheia (Clausner et al., 2011a) is used.

The process of generating the ground truth for the image is as follows:

Step One

The image is loaded into Aletheia and thresholded to produce a binary image, an example of which (thresholded using the Sauvola (2000) method) is shown below in :

Figure 4-9 Thresholded Image in Aletheia

Step Two

Then components that do not form part of the object(s) of interest are manually removed ().

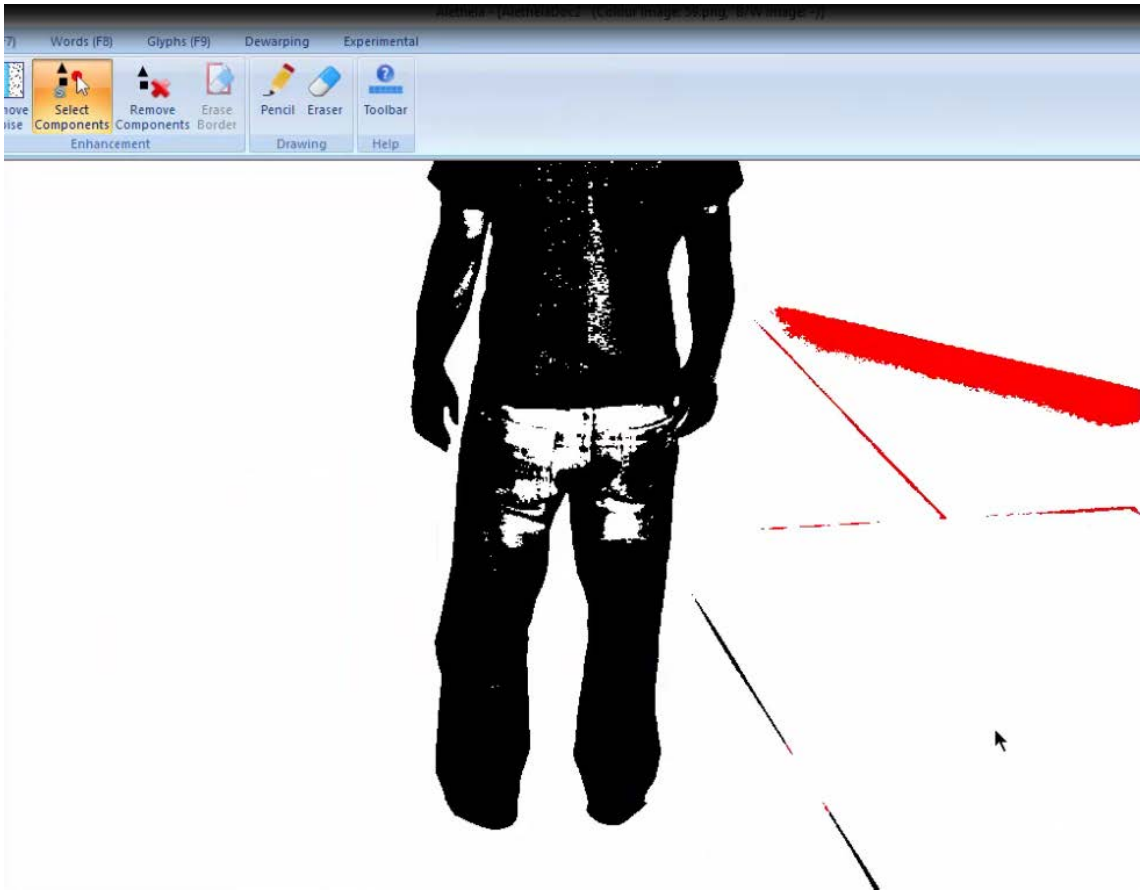


Figure 4-10 Manual Removal of Components

Step Three

Next, the silhouette border is identified: the area around the object is highlighted then a shrinking process fits the border to it ().

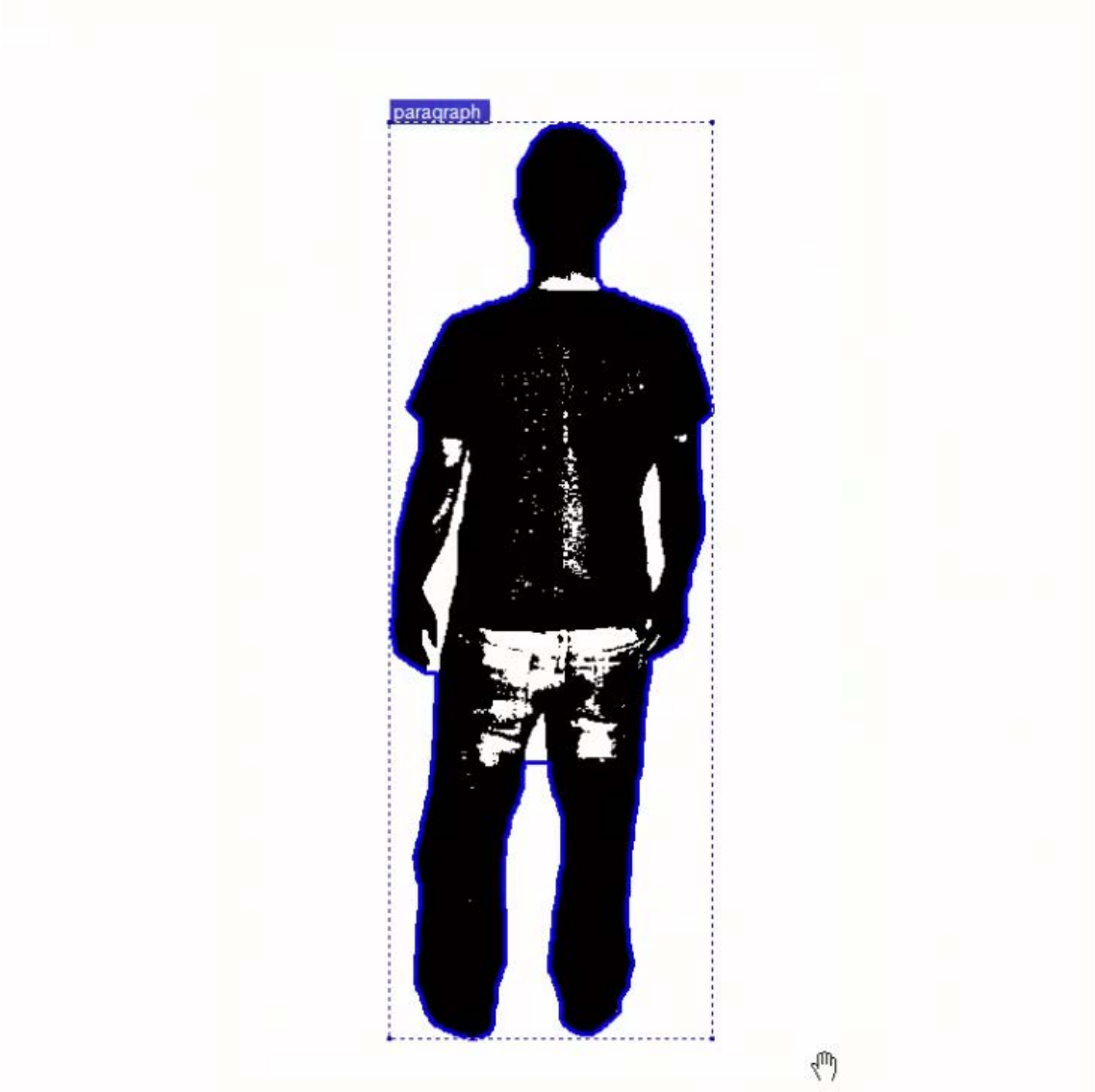


Figure 4-11 Silhouette Border Idenfication

Step Four

Finally the border is manually refined to fit the exact outline of the object ().

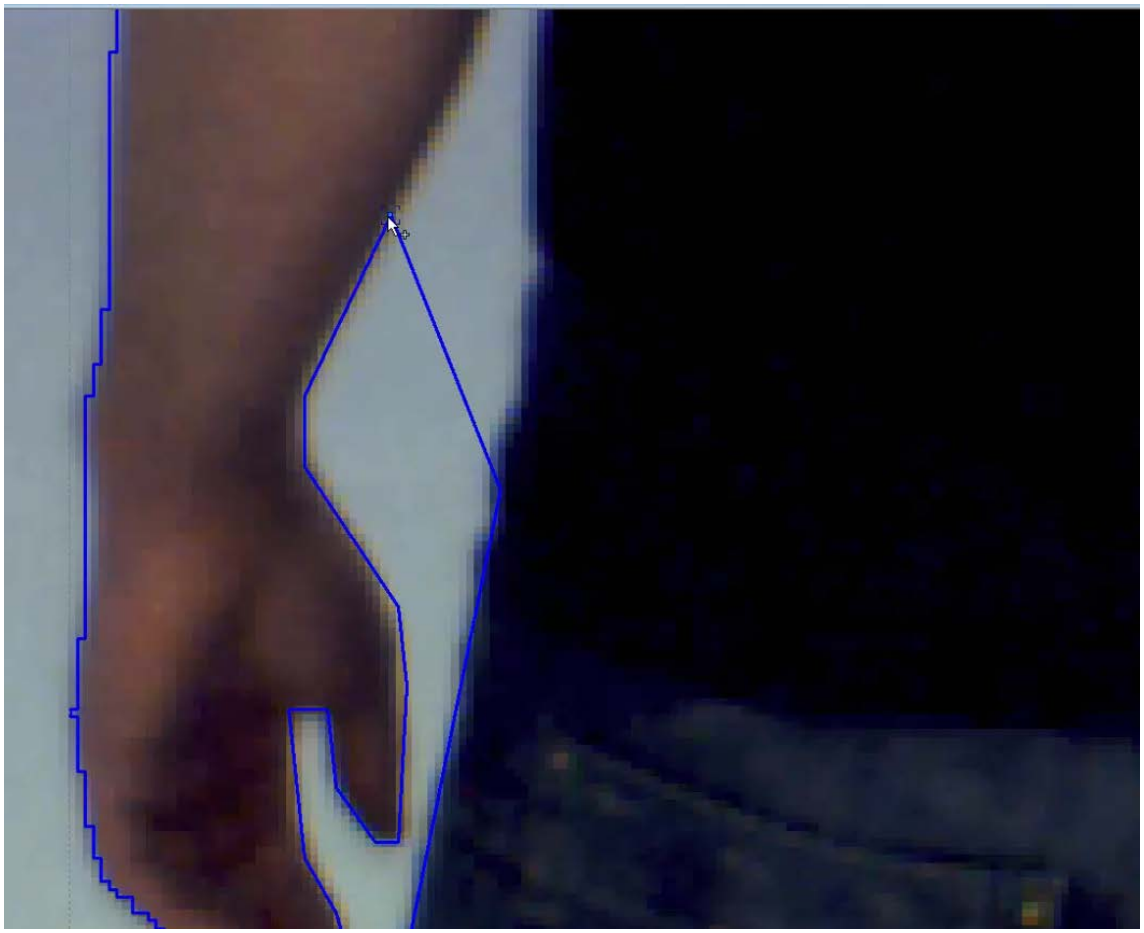


Figure 4-12 Silhouette Border Refinement

Once the ground truth is acquired methods of evaluation are determined. As an initial evaluation a comparison between the ground truth silhouette area and that of both Mester and MOG were made.

For area evaluation two different types of errors can be identified:

- The border of the silhouette falsely includes parts of the background area (INCL)
- The border of the silhouette falsely excludes parts of the foreground area (EXCL)

As shown in below:

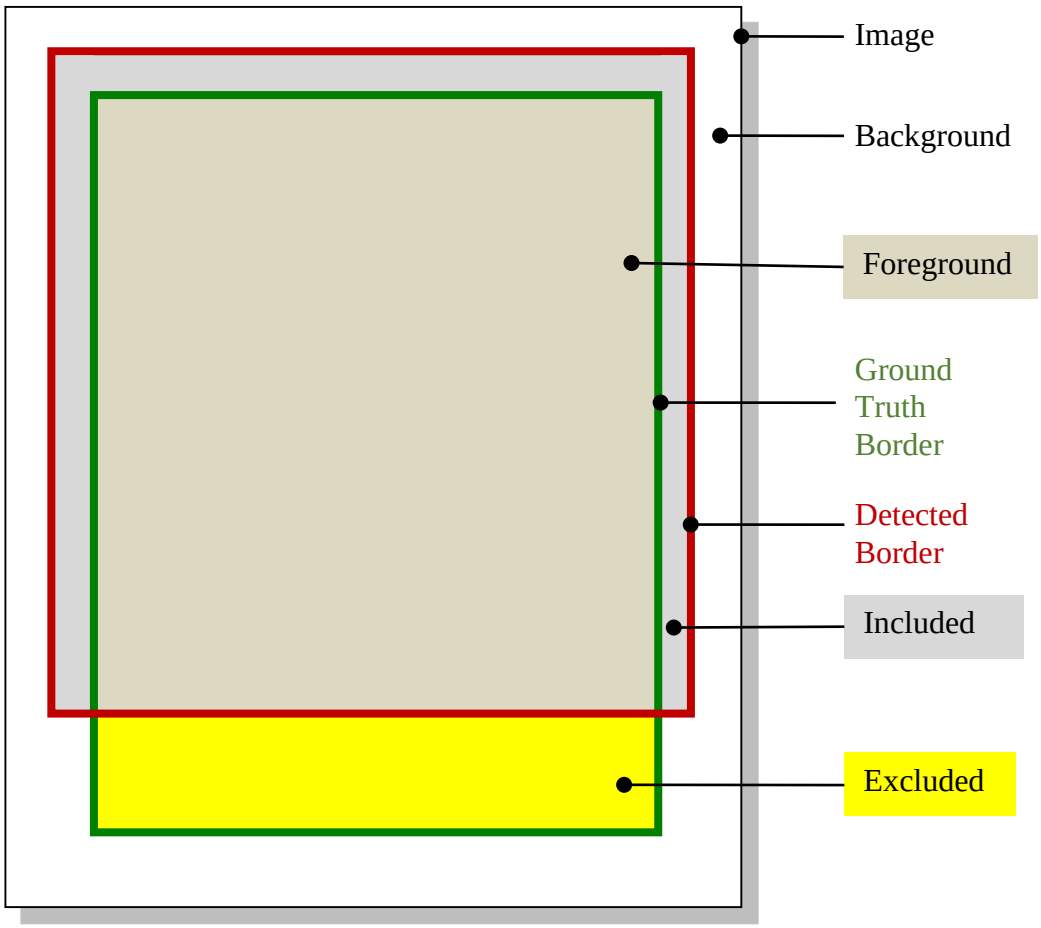


Figure 4-13 Example of Included and Excluded Areas

If GT is the area defined by the ground truth border and SEG the area defined by the automatically detected border, the error areas are calculated as follows:

$$INCL = SEG \setminus (GT \cap SEG)$$

$$EXCL = GT \setminus (GT \cap SEG)$$

Based on the error areas, weighted errors are calculated:

$$ERR_{INCL} = w_{INCL} * INCL$$

$$ERR_{EXCL} = w_{EXCL} * EXCL$$

Then, the success rates are calculated:

$$S_{INCL} = f_{INCL}(ERR_{INCL})$$

$$S_{EXCL} = f_{EXCL}(ERR_{EXCL})$$

The success rate functions are the nonlinear functions discussed in (Clausner et al., 2011b). The 50% points are set according to following table:

Table 4-1 Success Rate Settings

Type	50% Point
INCL	(IMG - GT) / 2
EXCL	GT / 2

Where IMG is the area of the whole image.

For the overall success, another function f_{IMG} with 50% point at (IMG/2) is used. The formula is:

$$S_{OVERALL} = 2 / (1/f_{IMG}(ERR_{INCL}) + 1/f_{IMG}(ERR_{EXCL}))$$ (harmonic mean)

GT performed to ten silhouette images.

The results from each image are detailed below:

Mester

Table 4-2 Success Rates for Mester

Ground-Truth	harmonicWeightedAreaSuccessRate
cam00.xml	99.5%
cam01.xml	99.8%
cam02.xml	99.8%
cam03.xml	99.7%
cam04.xml	99.7%
cam05.xml	99.6%
cam06.xml	99.8%
cam07.xml	99.7%
cam08.xml	99.7%
cam09.xml	99.7%
	99.70%

MOG

Table 4-3 Success Rates for MOG

Ground-Truth	harmonicWeightedAreaSuccessRate
cam00.xml	99.7%
cam01.xml	99.9%
cam02.xml	99.8%
cam03.xml	99.8%
cam04.xml	99.8%
cam05.xml	99.7%
cam06.xml	99.8%
cam07.xml	99.7%
cam08.xml	99.8%
cam09.xml	99.8%
	99.78%

The result of this type of evaluation shows that the MOG method is 0.08% more accurate that Mester for this sample set.

It should be noted that although small, this difference can be considered relevant since the measure is based on the area. It therefore takes into account the whole objects and not just the silhouette.

4.1.4.8 Performance

The implementation of the method is GPU based and requires features specific to versions of CUDA greater than the hardware initially available to the author but was later upgraded. The upgrade enabled the method to run but at a relatively low rate compared to what was achievable using faster hardware currently available (see).

Table 4-4 Segmentation Times for Processing Two Streams Simultaneously

Method	**GeForce 730**	**GeForce GTX 660**	**Quadro K5000**	**GeForce GTX 970**
MOG	27.17ms	5.79ms	5.31ms	2.82ms

4.1.5 Summary

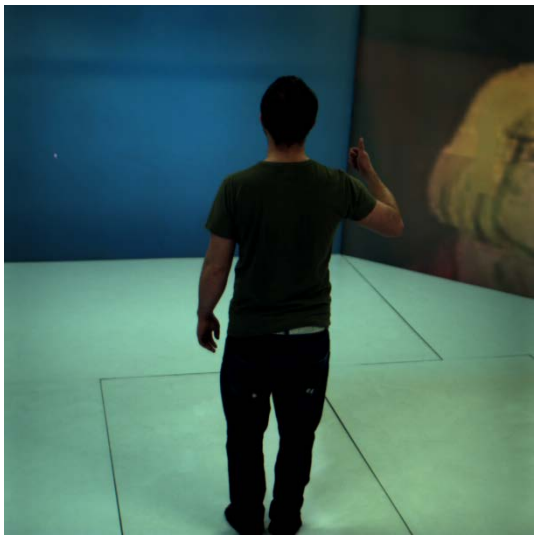
The requirements for background-foreground segmentation approach employed in the system were ease of use and accuracy of the result produced. It was clear that the Mester method fell short on both counts and a new approach to segmentation was required. The analysis of different methods available lead to the conclusion that MOG was the preferred choice and it results in consistently faithful 3D model reconstruction.

4.2 Infrared

Even with the improvements in Chapter 4.1 there is a major drawback inherent in this method of segmentation in visible light spectrum for telepresence applications.

Specifically, if an ICVE such as the Octaves display system was tuned on to allow a rendering of a user from another location or for anything else for that matter, this would interfere with the background subtraction segmentation that is currently in use and it is not possible to create a 3D avatar. There are approaches to overcome this detailed in chapter 3.8. The drawbacks are that only a partial 3D reconstruction can be created or a full 3D representation that is inferior and contains holes due to pattern interference.

The problem is best highlighted with an example of what happens to the current segmentation if a moving background is displayed on the projectors as in Figure 4-14.



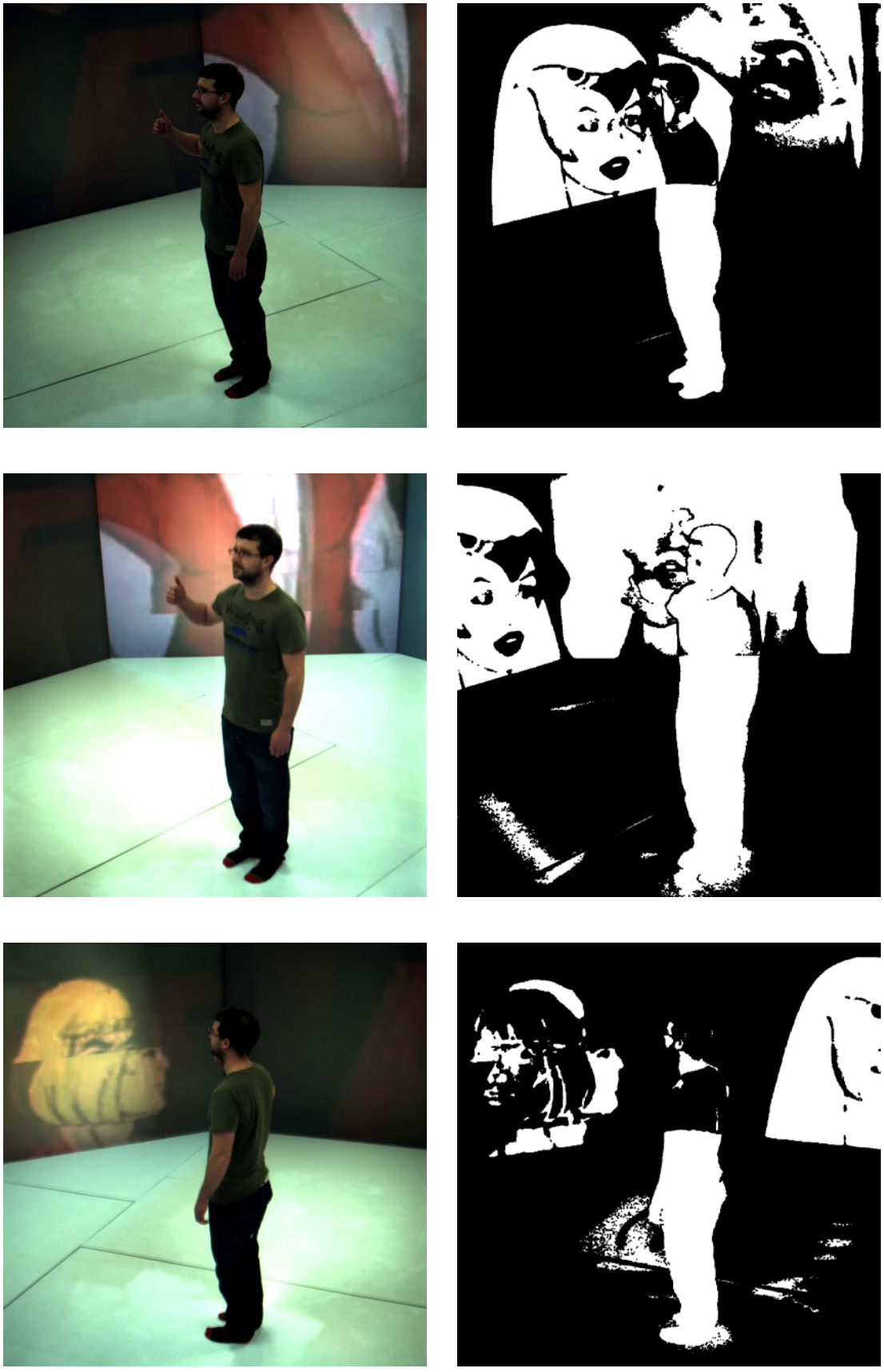


Figure 4-14 - Segmentation in Visible Light Spectrum whilst Video Projected

The segmentation employed is the MOG described in Chapter 4.1.4 and all images are acquired at the same time. It is evident that the current segmentation process is limited to sterile environments thus rendering it completely useless for a fully immersive IVT experience. In the past cameras have been positioned so one screen isn't in view so that at least one screen can be used for presentation. It is evident this is way off what is required to achieve the aim of IVT.

A novel segmentation process was proposed and implemented confining the segmentation to the Infrared spectrum. Because the projectors are not emitting light in the Infrared spectrum there output wouldn’t interfere with the segmentation process

An experimental rig was setup to test the hypothesis as shown below in Figure 4-15, Figure 4-16 and Figure 4-17:

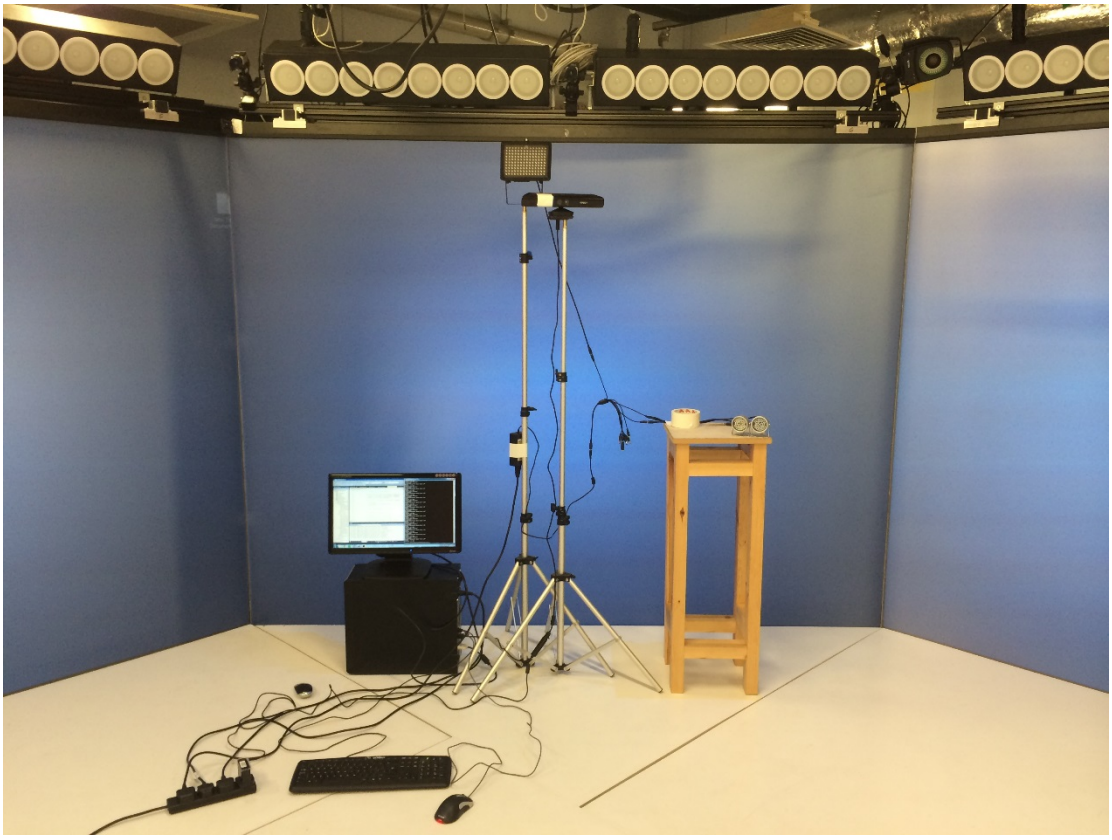


Figure 4-15 Infrared Segmentation Experimental Setup

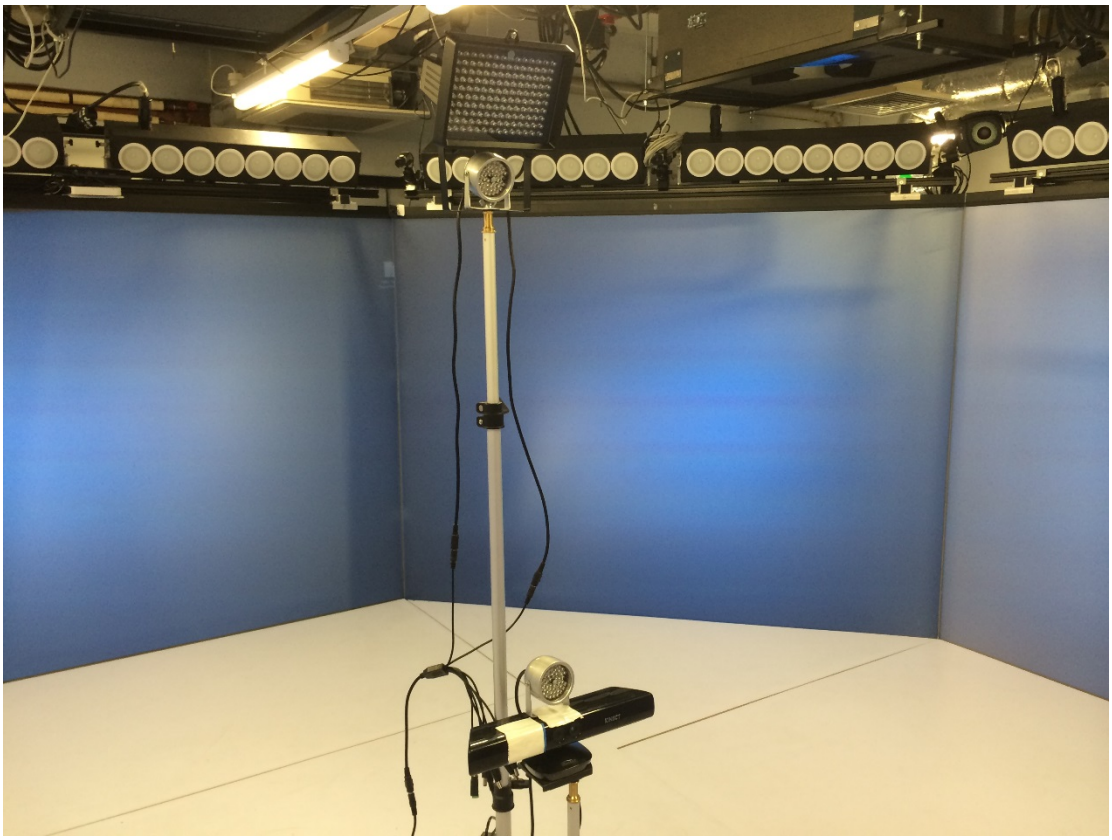


Figure 4-16 Infrared Segmentation Experimental Setup

Figure 4-17 Infrared Segmentation Experimental Setup

It comprises of a Kinect sensor, several Infrared lamps and a PC with the silhouette sender framework configured to utilise the Kinects Infrared camera.

It should be noted that the Kinect isn't being used as a depth-mapping device. The infrared camera is being utilised as it is an accessible commodity piece of hardware. The same process could be used with any camera capable of supporting an Infrared filter.

Results from the preliminary testing are shown below in Figure 4-18 and Figure 4-19:

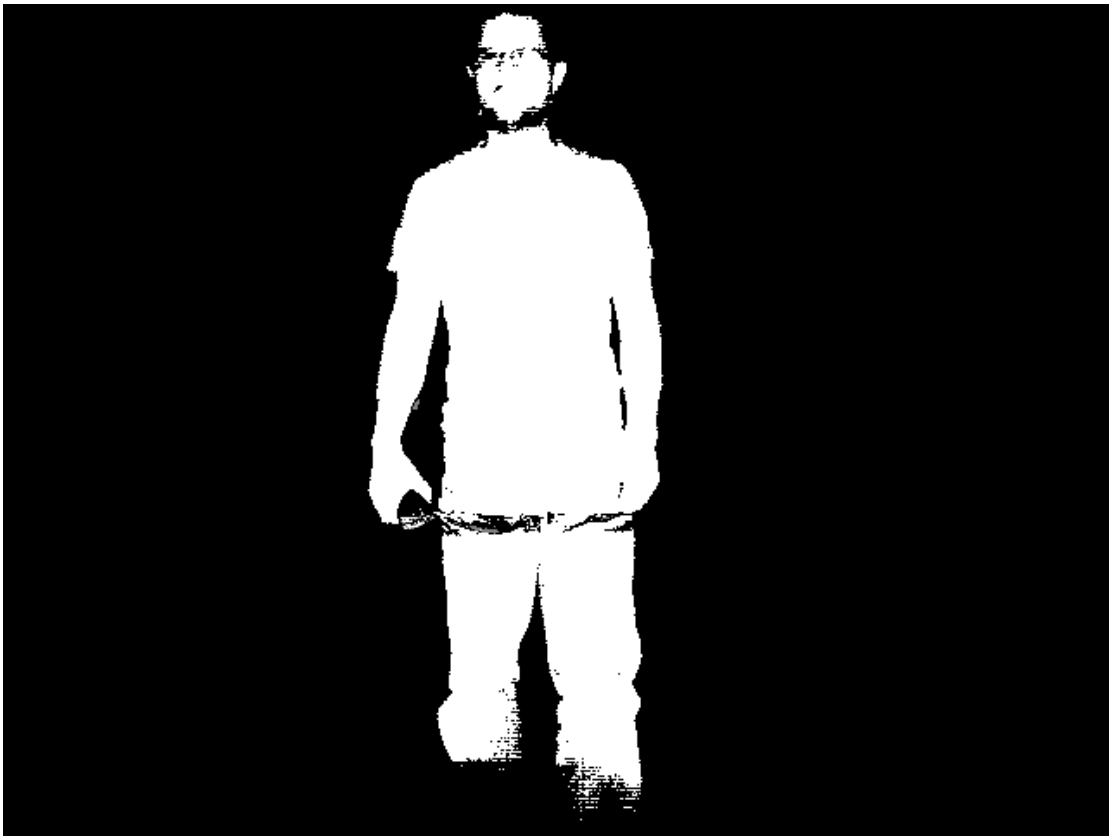


Figure 4-18 - Infrared Segmentation Result with Moving Background

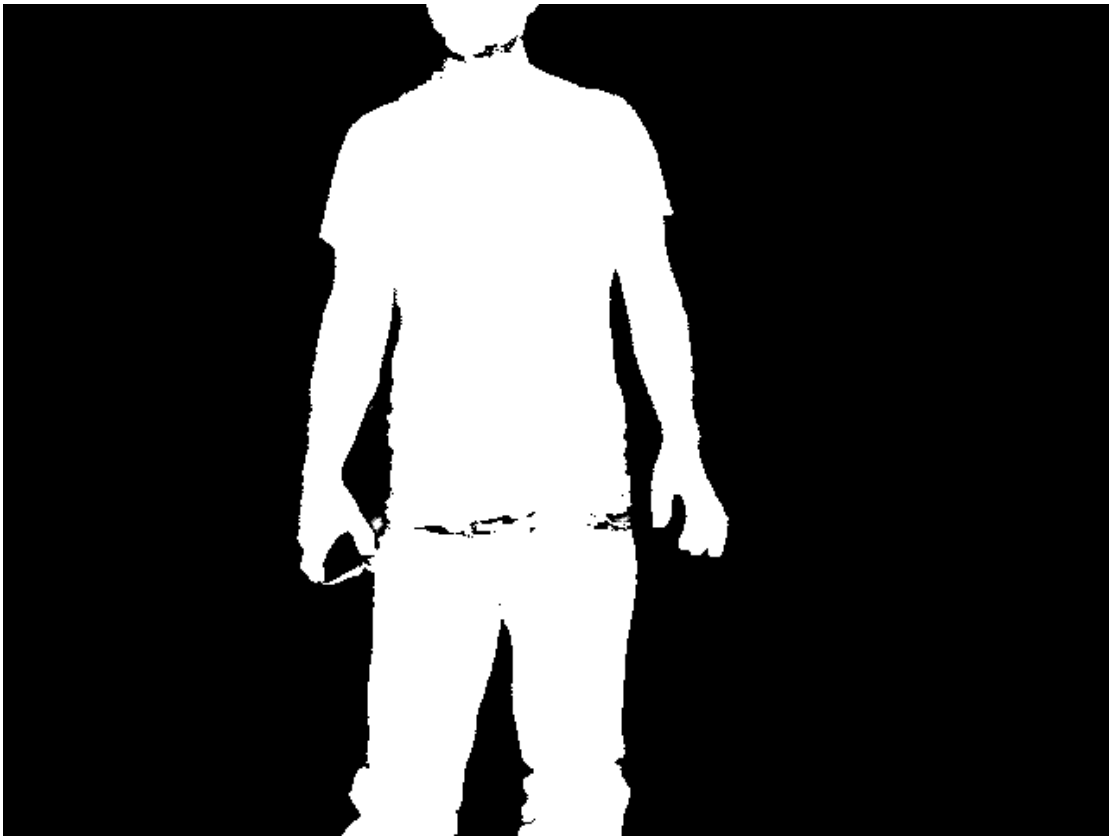


Figure 4-19 - Infrared Segmentation Result With Moving Background

They highlight the ability to successfully segment a user whilst live video is being projected on the surrounding monitor walls. They also highlight some issues with the proposed method: the infrared light sources used in the experiment were not able to illuminate the user with enough infrared light to demonstrate completely successful segmentation. The method of background segmentation used was the same as in 4.1.4. The problem is especially evident where the Infrared light is being absorbed rather than reflected in locations such as the hair and beard of the user.

Following on from the preliminary testing further IR light sources were sourced and different configuration of their pose tested in different environments. Here the results from further experimentation with additional light sources and various positions they

were configured in are presented. The additional lamps were tested in different configurations and with the user at varying distances from the lamps and cameras.

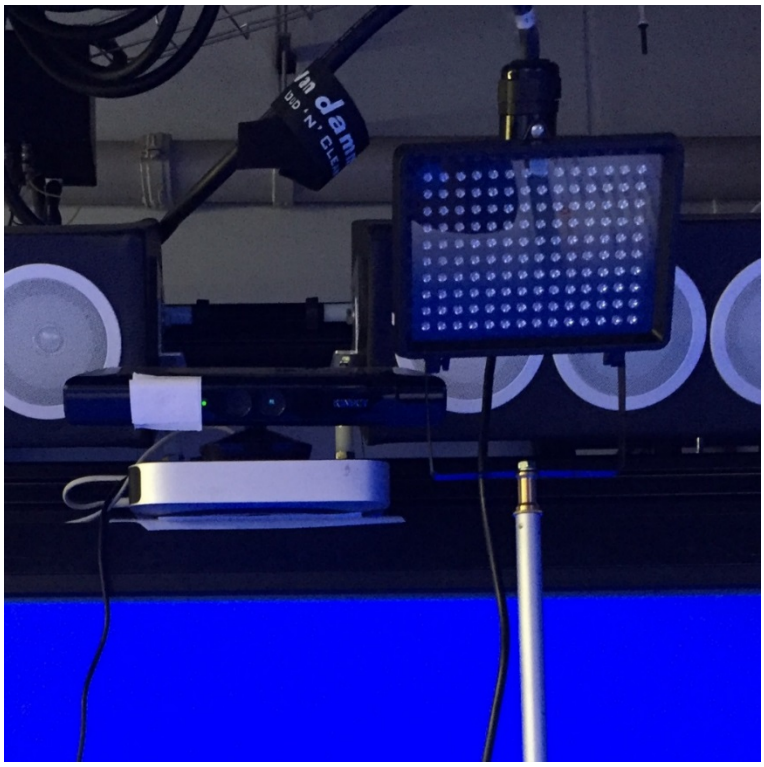


Figure 4-20 Infrared Lamp Positioned Alongside Kinect Sensor

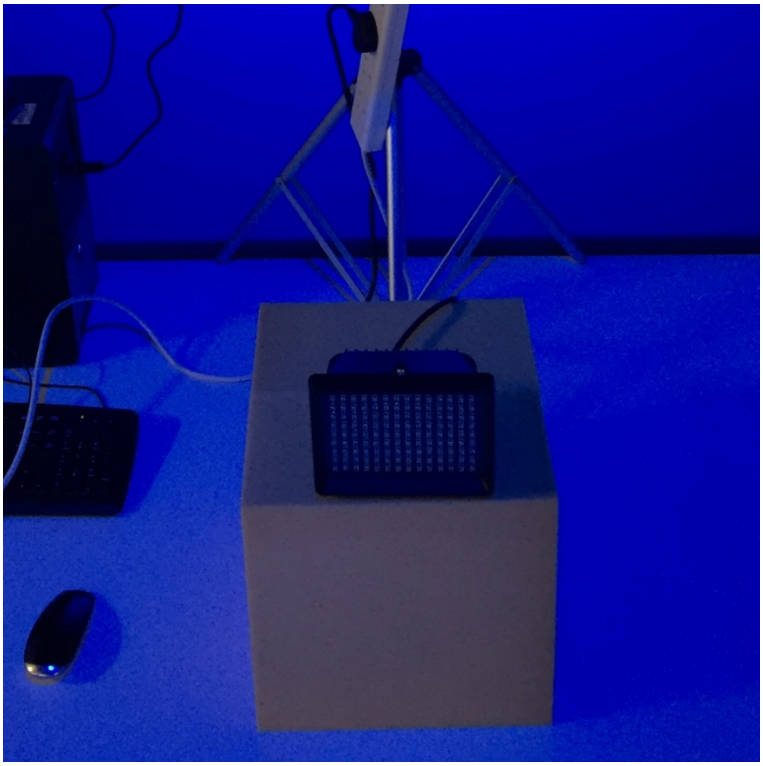


Figure 4-21 Infrared Lamp Positioned Slightly Above Floor Level

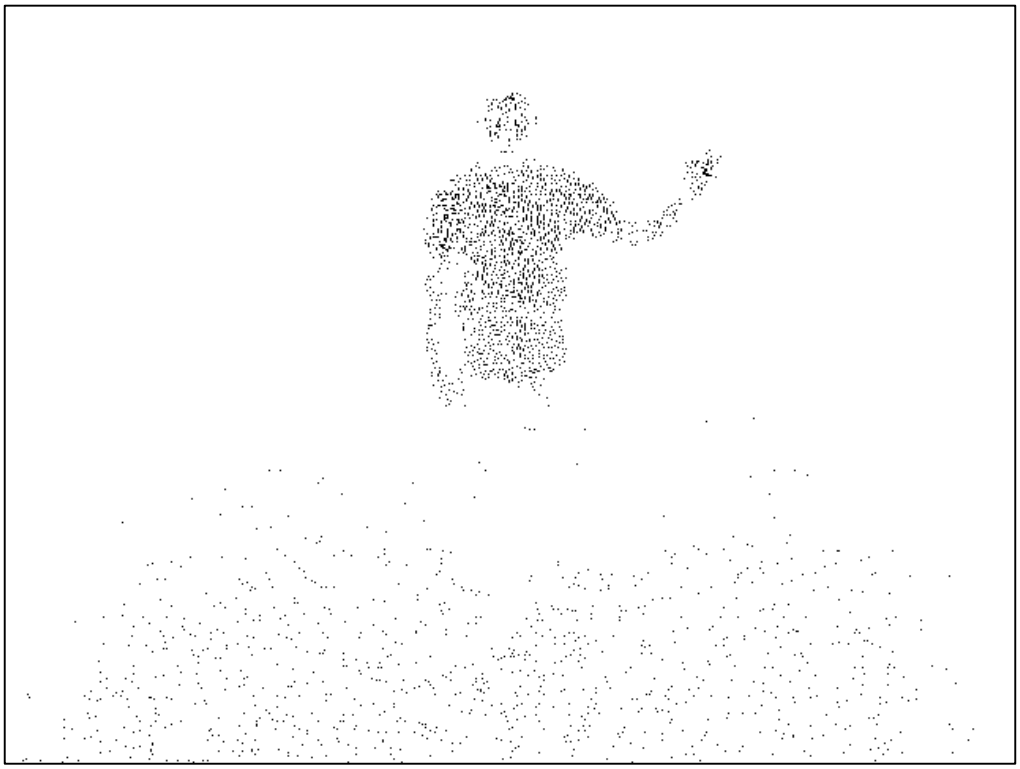


Figure 4-22 Segmentation Example with the Kinects Built-In Infrared Projector

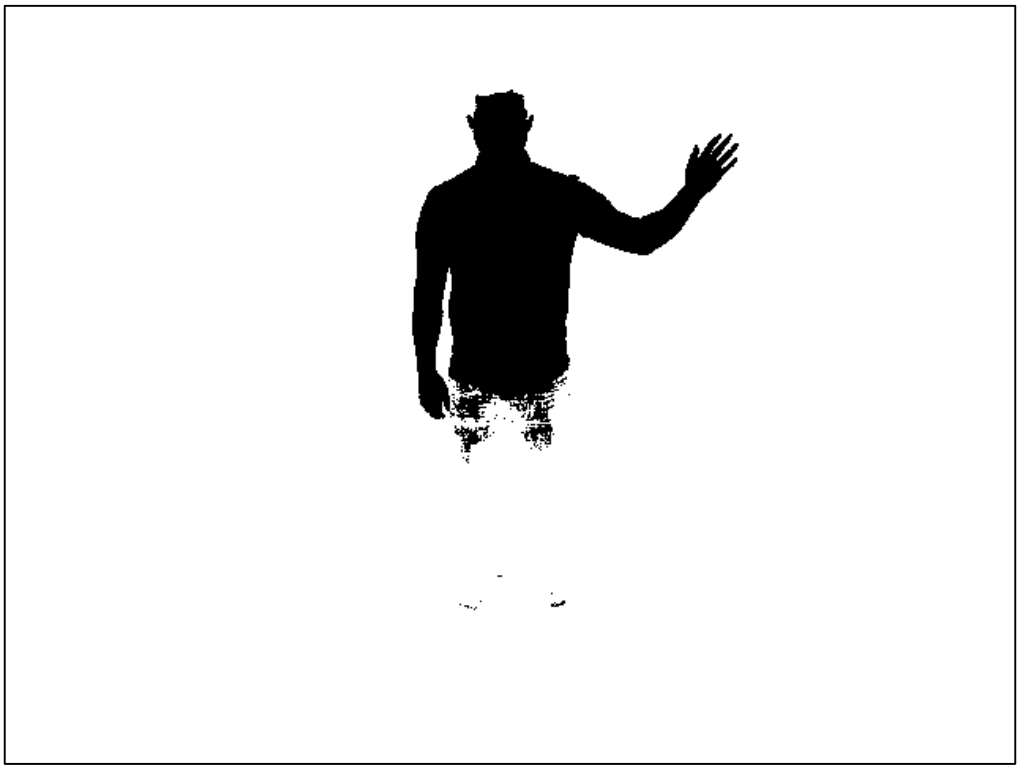


Figure 4-23 Segmentation Example with Infrared Lamp Next to Kinect

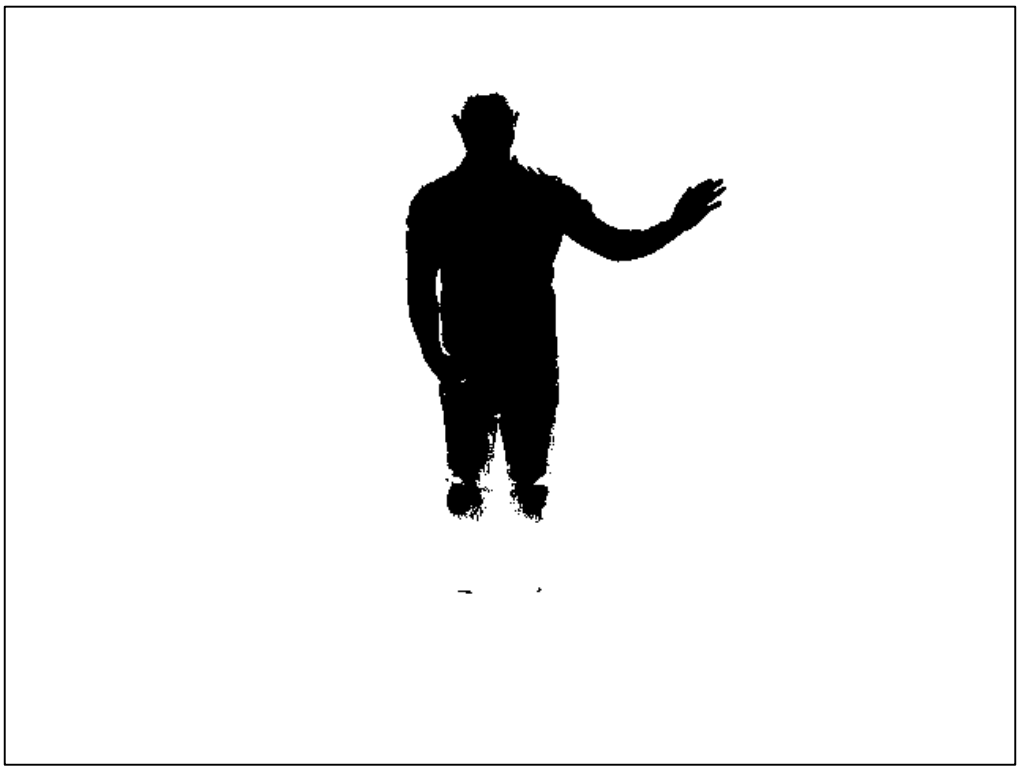


Figure 4-24 Segmentation Example with Infrared Lamp Slightly Above Floor Level

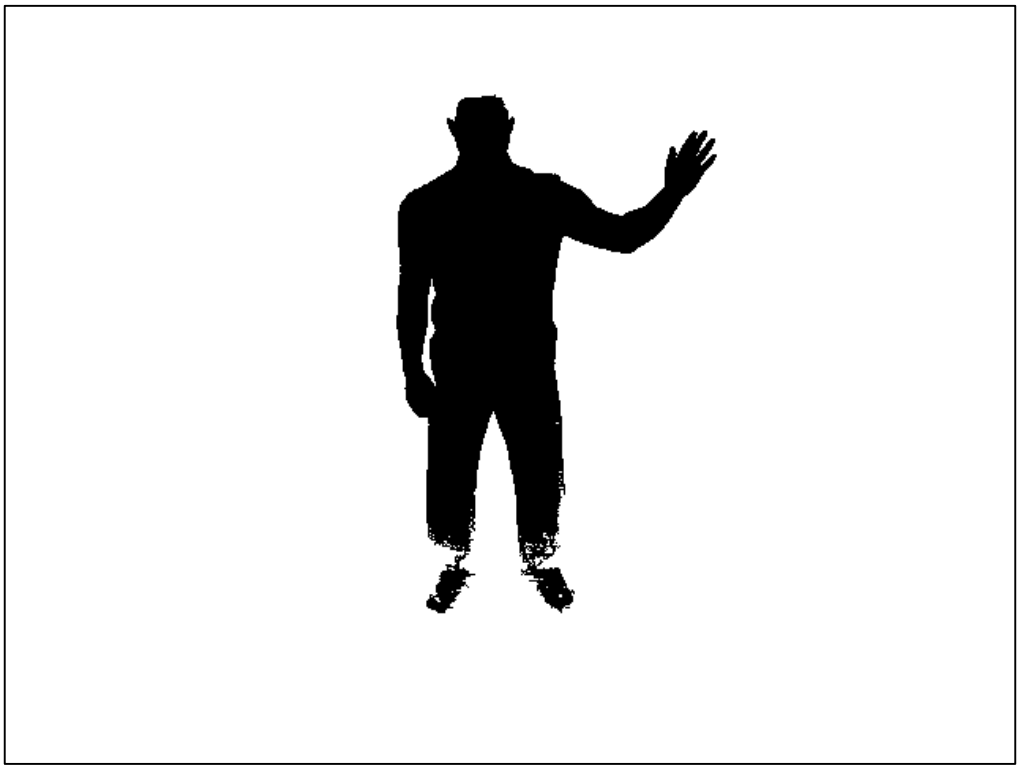


Figure 4-25 Segmentation Example Using Both Infrared Lamps

Distance from the Infrared light sources and Kinect sensor arrangement.

Figure 4-26 Distance A

Figure 4-27 Distance B

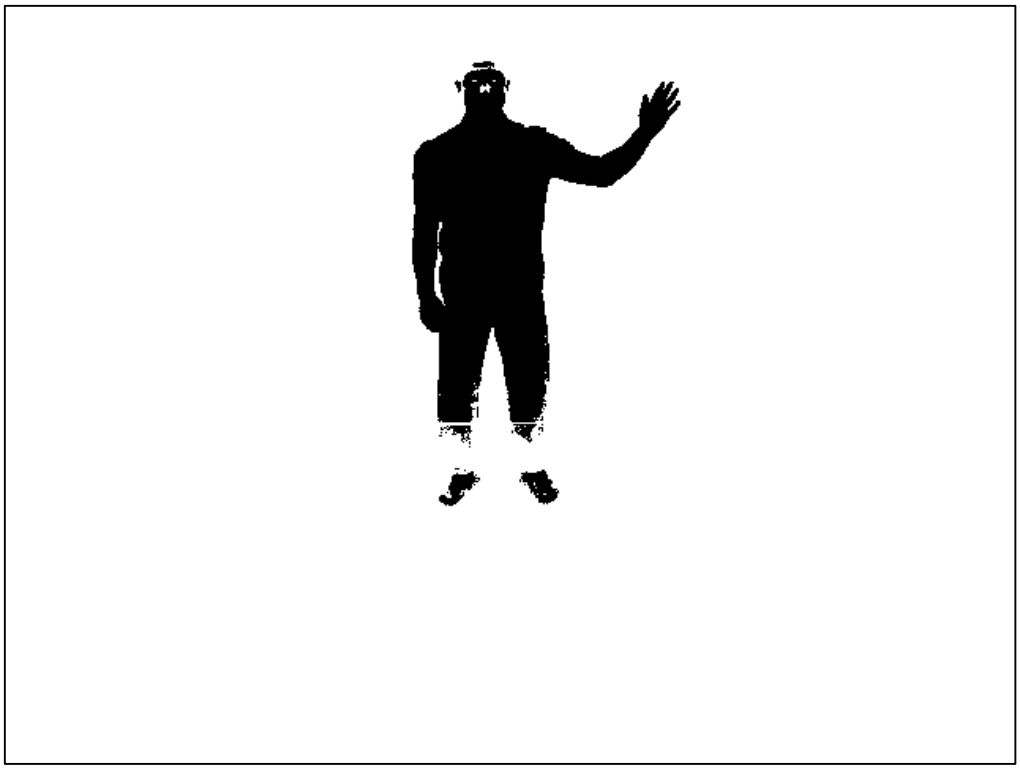


Figure 4-28 Distance C

The testing was also conducted at a different location to the Octave namely ThinkLab to determine if the same properties of the setup held true:

Figure 4-29 ThinkLab

Figure 4-30 ThinkLab Infrared Setup

Figure 4-31 ThinkLab Infrared Result

The results show that the IR light source placement and power output affects the amount of information available in the images and thus the quality of segmentation. Further testing wasn't possible in the time scale of the research but suggestions are made in the discussion.

4.2.1 Summary

The solution presented demonstrates the potential to capture silhouettes whilst live video is presented to the participating user. The results section presents further testing with additional Infrared light sources surrounding the user. The result is a resolution of 640 x 480 pixels, which surpasses the current resolution available with the Kinect depth mapping of 320 x 240 pixels.

Chapter 5

System Architecture

In this chapter the end-to-end system architecture is presented. The process begins with the acquisition then segmentation of the subject(s) and object(s) of interest, followed by the 3D reconstruction, then the distribution and rendering. First, the end-to-end system architecture is detailed followed by a description of each stage in the pipeline. Also improvements over previous architecture and the components that have been retained are detailed.

5.1 System Architecture

The complete end-to-end system architecture is comprised of multiple network-connected components with each contributing to the processing pipeline that is described below and visualized in Figure 5-1. This figure depicts the high level visualization of the system architecture from the cameras that acquire the subject(s) and object(s) (top) through to rendering on the end nodes, which can be in different and possibly geographically dispersed locations (bottom). The following subsections outline the end-to-end process.

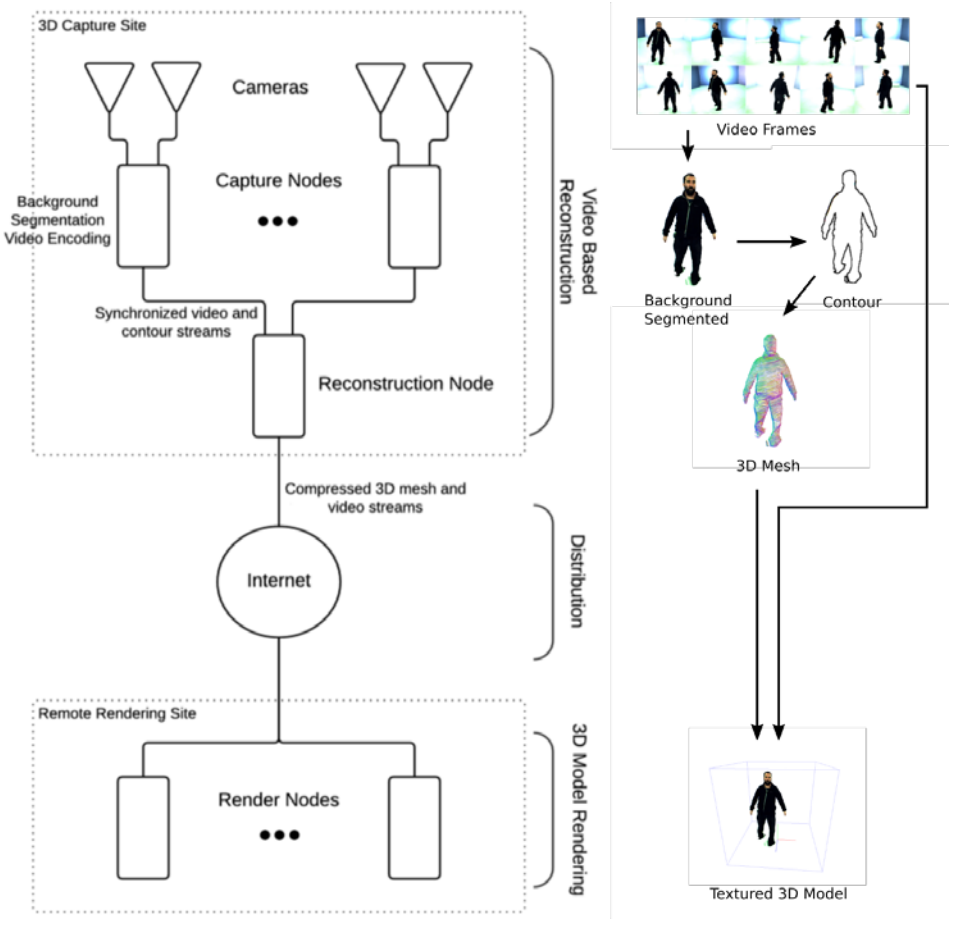


Figure 5-1 Generic System Architecture

5.1.1 Subject Acquisition

The process begins with the acquisition of the subject(s) via an array of cameras surrounding and directed towards them in the capture volume. The capture volume could be setup in a room such as an office but often the Octave is utilised. Cameras are either mounted on tripods or above the displays depending on the display configuration. A panoramic image of the Octave is shown in Figure 5-2 with a UML Deployment Diagram shown in Figure 5-3.

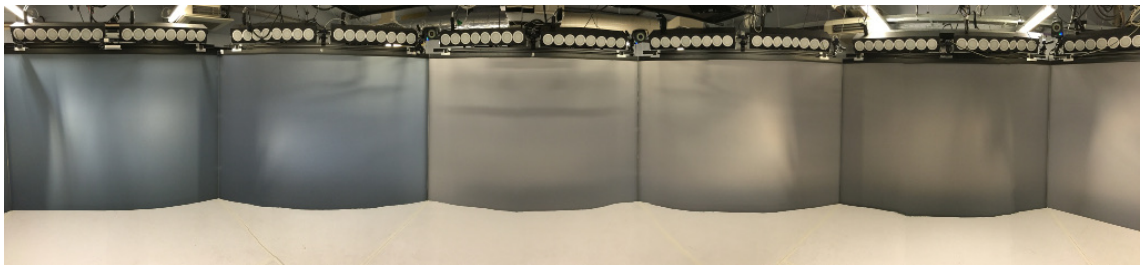


Figure 5-2 Panoramic Image of the Octave

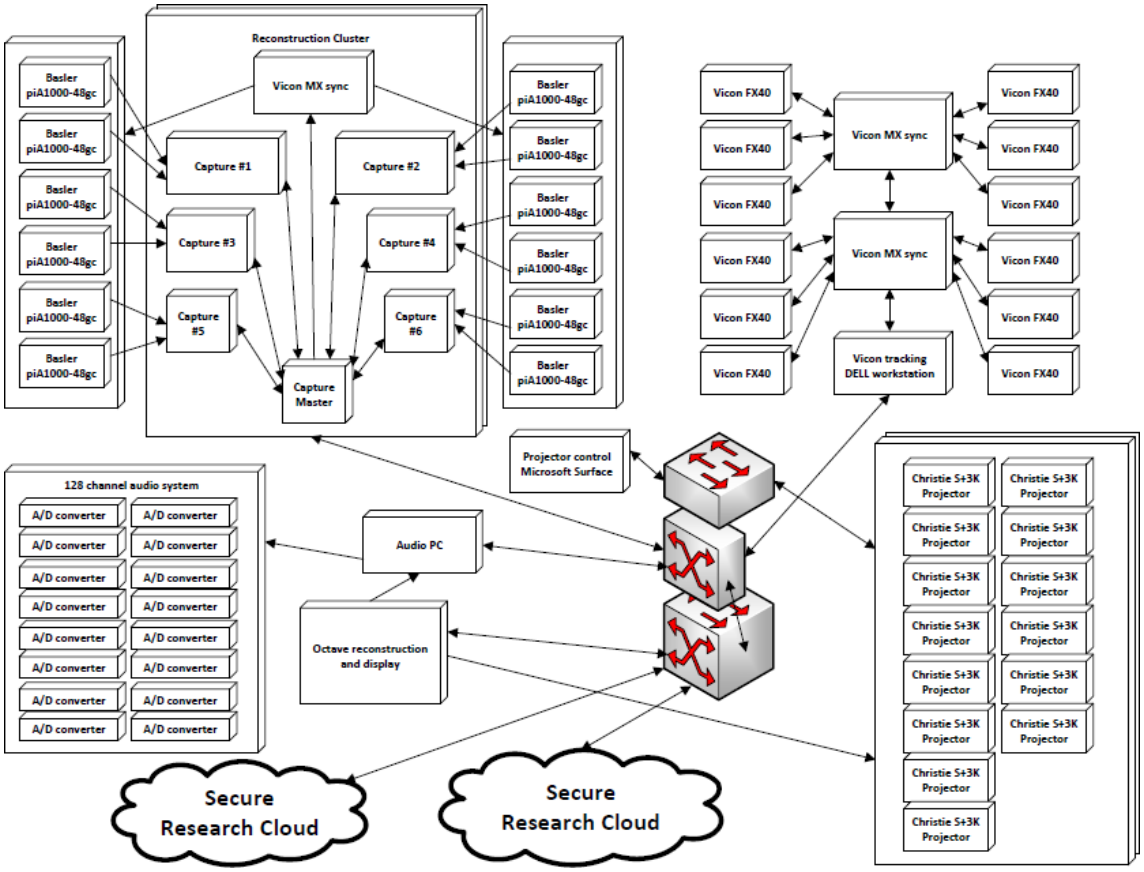


Figure 5-3 UML Deployment Diagram of the Octave

The 3D reconstruction process requires the acquisition of images and their corresponding silhouettes to be temporally synchronised across all cameras in order to produce faithful results. The prototype developed to investigate the distribution and processing of video for real-time 3D telepresence (Moore, 2012) demonstrated its capability to acquire synchronous images from cameras, encode to compressed video streams then distribute to the reconstruction node. It was evaluated in the context of the requirements for the system being developed to determine if any components could be retained and utilised. In the prototype, the cameras are connected to capture nodes, which are commodity PCs running the bespoke capture software. They are connected to the same switch on which the reconstruction node resides and it utilises an exposed method in the capture software

to request frames from them. The UML Deployment Diagram shown below depicts the complete end-to-end system.

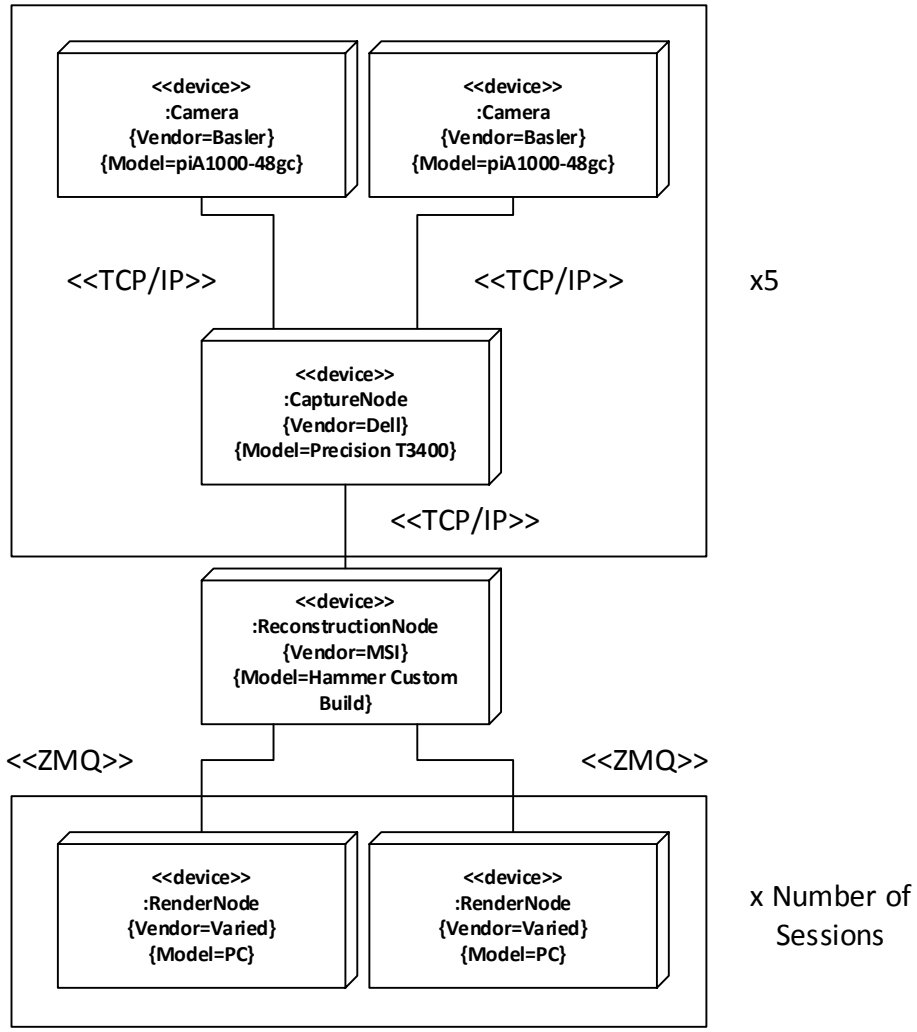


Figure 5-4 End-to-End System UML Deployment Diagram

5.1.2 Capture Nodes

The capture nodes facilitate the synchronous acquisition of frames from the cameras, lens distortion correction (if necessary), foreground-background segmentation and the encoding of the images in to compressed video streams.

5.1.2.1 Synchronous Frame Acquisition

An evaluation of the prototype determined that it could:

1. Acquire frames from the cameras,
2. Encode these frames as compressed video,
3. Transmit and decode the video back into frames on the reconstruction node.

Using ten cameras (with a resolution of 1000 x 1000 pixels) connected in pairs to five capture nodes it was capable of a sustained capture rate of 25 frames per second. This confirmed that the prototype had real time potential.

However, during the evaluation a number of shortcomings were identified that affected the suitability of the prototype for use in this system:

The capture across all the cameras typically used in the experimental setup was not temporally synchronised to the accuracy required. It was determined that the reason for this was the method of acquiring images from the cameras and not with the software on the capture nodes. To remedy this issue the documentation for the cameras programming interface was consulted and a more robust method to request frames was discovered and implemented: rather than continuously requesting frames and simply using the latest contained in a buffer a software trigger was implemented which instructs the camera to acquire an image at a distinct time. The use of this method resolved the issue without any effect on framerate.

The lead node (which is the one running the 3D Reconstruction process) was configured to pull a video and silhouette sequence from the other nodes once it had completed the reconstruction process. This adds latency to the system and with this in mind a completely separate thread and buffer system was added to the lead node software to decouple the image acquisition from reconstruction depicted in Figure 5-5.

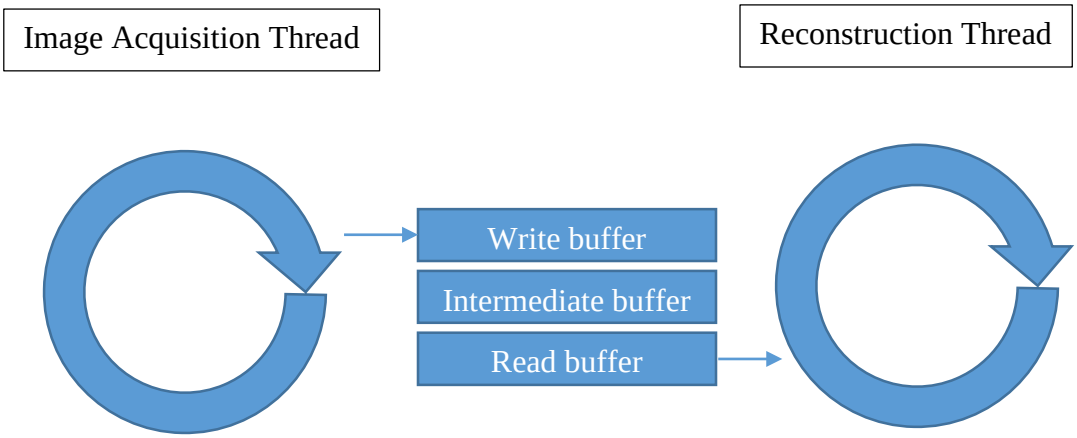


Figure 5-5 Image Acquisition and Reconstruction Threads and Buffer

5.1.2.2 Lens Distortion Correction

This process takes place on the capture node post frame acquisition and can be invoked on a per camera basis if required. It uses an implementation of Chessboard calibration (Z. Zhang, 2000) from the OpenCV library and is executed on the capture nodes. Generally, the correction process is only necessary when using wide-angle lenses or those of lesser quality.

5.1.2.3 Colour Correction

The process of correcting the colour across the cameras was given much thought and attention as detailed in Chapter 3.2. In this chapter an advanced method to correct colour is presented but it was not implemented into the system. Instead, a simpler approach to allow the correcting the brightness and contrast was used. In this implementation, if required, the brightness and contrast of input images can be altered post capture in order to correct the image if necessary. In keeping with a straightforward system design the brightness and contrast settings to be applied are integrated with the camera settings

configuration, which is stored in XML format on the 3D reconstruction node. The settings are distributed to all the capture nodes on initialisation.

The process is only invoked if the contrast setting is not equal to one or the brightness is not equal to zero.

Iterates through all pixels performing the following alteration (where *alpha* is the contrast and *beta* is the brightness):

```
m_frame32Mat->at<cv::Vec4b>(y,x)[c] = cv::saturate_cast<uchar>( alpha*(
m_frame32Mat->at<cv::Vec4b>(y,x)[c] ) + beta );
```

5.1.2.4 Background-Foreground Segmentation

The process is executed on the capture node post-lens distortion and colour correction. The process results in a silhouette of the foreground. Ideally this foreground is just the objects that the 3D form is intended to be reconstructed from. This accuracy of the result of this process is of crucial importance and directly linked to the faithfulness of the 3D hull reconstruction process and therefore the final resultant avatar(s). It is discussed in Chapter 0.

5.1.2.5 Video Encoding

Transmitting multiple streams of uncompressed frames to the reconstruction node is not achievable given current the bandwidth constraints in current commodity hardware. The capture nodes navigate this constraint by encoding the video into compressed format before transmission. Currently H.264 video compression is used. It was determined that there was no perceivable reduction in image quality using this method once the encoding had gained enough entropy. Experimentation determined that this typical takes around 100 frames.

5.1.3 Capture Node Usability Enhancements

In keeping with the overarching aim of a usable system it is worth mentioning some key usability enhancements that were made.

5.1.3.1 Changing Parameters and Settings

Previously, changing the camera settings was a time consuming process requiring specific knowledge of the camera software to be applied on each capture node. Considering that the settings required changing for calibration and experimentation the ability to change the camera settings across all nodes from a single node was added. Furthermore, the camera settings are contained in an XML file so different configurations can be saved and loaded accordingly. The XML syntax for the camera settings are shown in the following table:

Table 5-1 XML Syntax for Camera Settings

Element	Property
<CameraSettings></CameraSettings>	Main XML element specifying that everything enclosed is camera settings.
<CameraSetting />	XML element for each camera with attributes for: **id** *the id of the camera* **red**, **green**, **blue**, **exposure**, **blackLevel**, **gain** *settings applied directly via the cameras API* **contrast, brightness** *settings applied post acquisition on the camera node*

For example:

```
<CameraSettings>
<CameraSetting id="0" gain="100" red="60" green="80" blue="140" exposure="800"
blackLevel="10" contrast="1" brightness="0" />
<CameraSetting id="1" gain="100" red="60" green="80" blue="140" exposure="800"
blackLevel="10" contrast="1" brightness="0" />
…
</CameraSettings>
```

5.1.3.2 Starting and Stopping the Capture Nodes

A series of batch processor scripts were created to aid users with the capture system start-up and shutdown. To further enhance the usability of the system the batch processor scripts employ the use of a remote execution tool that allows the start-up of the capture software on all nodes from a single location. To remove any ambiguity regarding settings the scripts are configured to use the appropriate XML files.

5.1.3.3 Calibration Mode

A new mode of operation was added to the capture software that starts it with a configuration specific to calibrating. The mode can be toggled in an XML configuration file on the reconstruction node with all other nodes being informed of the mode during their initialisation. When the calibration mode is invoked the capture nodes switch from their default mode of perform image acquisition, background-foreground segmentation and video encoding to sphere coordinate location (see Chapter 3.1.2.2). This is a very useful feature and makes the system much more useable. The XML syntax for the configuration of the reconstruction node is as follows:

Table 5-2 XML Syntax for the Reconstruction Node

Element	Property
`<CentralServer></CentralServer>`	Main XML element specifying that everything enclosed is central server settings.
`<Capture />`	XML element containing the capture type to invoke on the capture nodes with attributes for: **device** *the type of device: Pylon (Basler camera), Kinect, Webcam*
`<BroadcastIP />`	XML element used to denote the IP address of the lead capture node that will broadcast UDP packets so the other nodes will acquire an image synchronously. With attributes for: **ip** *the ip address*

	port *the port specified in the broadcast*
<ClientIP />	Denotes the IP addresses of each capture node present in the network. With the following attributes: **ip** *the ip address* **port** *the port the node will listen to for commands* **isRelay** *denotes if the node is the broadcast node with a true or false string*

5.1.3.4 Further capture devices added

In addition to the previous changes the capture node software was updated to acquire frames from both the visible light camera and Infrared camera in the Kinect sensor.

5.1.4 3D Model Generation

Upon receiving and decoding the video streams and contour data from the capture nodes, the reconstruction node generates a 3D model avatar via a parallelized EPVH implementation (Tobias Duckworth & Roberts, 2014), specifically the multiple CPU version which proved the fastest. To generate a 3D model the system requires knowledge of the cameras image planes in relation to real word 3D coordinates (Tsai, 1987). The accuracy of the result from the calibration process is of vital importance and directly related to the faithfulness of the reconstruction. It was the focus of much attention and discussed in detail in Chapter 3.1.

5.1.5 Distribution

After the 3D model generation, the reconstruction node prepares the 3D mesh and video data for distribution to connected remote rendering sites, packaging all relevant data into a network message.

The method of distributing the message to the render nodes is ZeroMQ (2015). This facilitates the transport of the message across any Internet Protocol network. The reconstruction node is the server, which accepts incoming connections from the render client(s). It creates a listening daemon on a TCP/IP port specified on start-up via a command line option.

The port listens for and accepts connections from the client render nodes. Once connected the server begins sending the clients messages which are serialised using Protocol Buffers (2015).

There are two types of message: calibration and 3D avatar data, which are detailed in the following two subsections:

5.1.5.1 Calibration Message

The rendering sites also require calibration data so they can correctly position cameras for the texturing process. As this calibration data is likely to change over time the system distributes it at the start of a session and at intervals of one minute for any late joining participants. This removes any requirement for loading the calibration at the remote sites. The format of the message is as follows:

CalibrationParamData	…	CalibrationParamData

Figure 5-6 Calibration Message Format

Where each CalibrationParamData message contains the following serialised data:

```
cameraID
focal_length_x, focal_length_y
r11, r12, r13
r21, r22, r23
r31, r32, r33
translation_x, translation_y, translation_z
```

Another advantage of storing the calibration data within the message stream is for post capture analysis. Previously the correct calibration file was required alongside the message stream that was stored to disk. Integrating the calibration with the message removes any ambiguity as to which calibration file belongs with a message stream thus providing another benefit in terms of usability.

5.1.5.2 3D Avatar Message

A message contains vertex positions, triangle indices, a video frame per camera, as well as frame number and timestamp (). In order to reduce the amount of data sent across the network, the 3D mesh data is compressed using the LZMA algorithm (7-ZIP, 2015) after serialization, and this results in between 67% and 75% reduction in size.

Frame no.	Timestamp	Compressed vertices	Compressed indices	Video data

Figure 5-7 Network Message Format

The video stream data component of the message is formatted as follows:

No. of streams	Stream length (in bytes)	Stream data	Stream length (in bytes)	Stream data	…

Figure 5-8 Video Stream Message Format

The video stream data is parsed as follows: first number of streams contained are read. Then, the stream length for the first stream is read and the subsequent bytes contained are copied and then decoded via the corresponding video decoder instance. Subsequent data is read in the same manner until the total number of streams have been successfully parsed.

The H.264 encoded video is taken directly from the input of the capture nodes to avoid decompression and recompression by the reconstruction component. Synchronization of

the video and mesh data is handled by placing the data together in the same network message.

5.1.5.3 Saving to Disk

In addition to live streaming a requirement of the system was for it to support the recording of sessions and playback in natural time. In order to fulfil this requirement, the Protocol Buffer messages can be saved to disk. Saving can be done instead of or in addition to live streaming.

5.1.6 3D Model Rendering

The Render Node Client performs the 3D Model Rendering and it has two modes of operation: live streaming and offline parsing from disk. The code base was designed to be reusable and has been integrated in to a proof-of-concept prototype.

5.1.6.1 Live Streaming

In live streaming mode the Render Node connects to the Reconstruction Node via a TCP/IP connection to a single port listening on a single IP address.

The IP address and port are specified via a command line argument.

Each message received via the TCP/IP is then processed as detailed in Subsection 1.

5.1.6.2 Offline Parsing from Disk

In this mode of operation the Render Node Client parses messages that have previously been saved to disk.

To ensure that the 3D model rendering plays back in natural time the loading is decoupled from the rendering by means of a separate thread in the client. This threads purpose is

time management. It loads messages but only signals to the rendering thread that a new message is ready once the time contained in its time stamp has elapsed. It is depicted in .

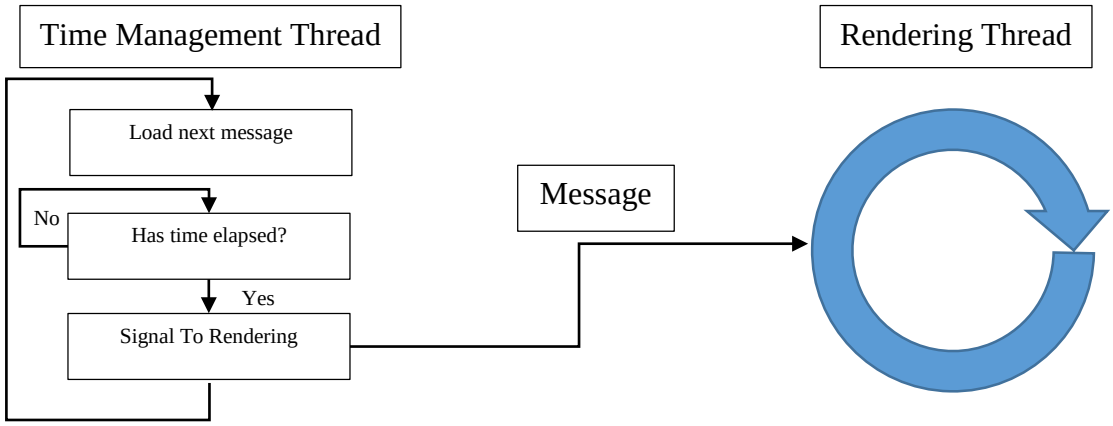


Figure 5-9 Time Management and Rendering Threads

The message is then parsed and rendered as detailed below:

5.1.6.3 Message Parsing

Upon receiving the message via the TCP/IP connection or disk parser it is de-serialised. If the message contains calibration information and the render hasn't had its calibration initialised from a previous message, then this process is invoked and the cameras are placed within the scene.

If the message contains 3D avatar data, then the compressed mesh and encoded video components are copied into respective buffers. The mesh is decompressed and the encoded video data is parsed with each cameras stream being directed to its corresponding decoder to prepare the next frame. Then rendering and texturing process then follows.

5.1.6.4 Rendering and Texturing Process

The render node computes vertex normals via the weighted average of the angle between connected triangle edges. It then pushes the vertex positions, normals and triangle indices

into OpenGL buffers on the GPU. The compressed video frames are decoded and pushed directly onto texture buffers on the GPU.

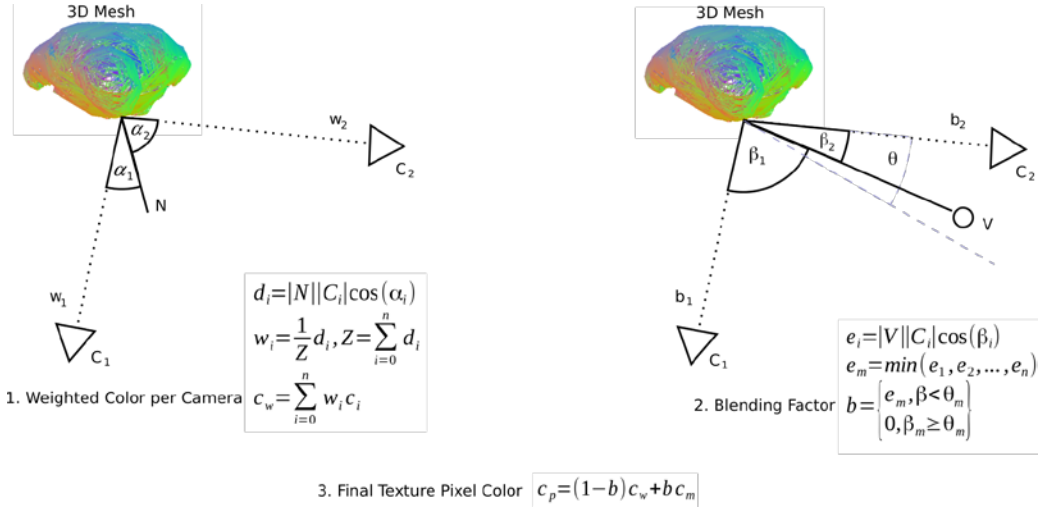
Texturing is realized in a pixel shader program in which each texture is projected onto the mesh from the corresponding camera perspective. The algorithm computes the colour of a pixel based on a weighted blending of projected pixels from the camera images. The blending weights w are determined by computing the dot product between the fragment normal N and the direction from the fragment to a camera C, so that $w = 1$ with $\alpha = 0°$ and $w = 0$ with $\alpha \geq 90°$. The weights are then normalized so that their sum equals one and applied when adding the projected pixel colours of the respective camera images, see Fig Figure 5-10, 1.

The weighted blending method provides that surfaces facing closer toward a specific camera receive a higher contribution to the final pixel colour from this camera's image than from others. The result is a smooth blending of the projected textures. While this method is simple and does not require on a specific camera arrangement, it can cause distortions in areas without a dominant camera and where cameras have similar weights. Furthermore, it does not take occluded areas into account.

In order to further improve visual quality, the texture mapping algorithm has been extended with a viewpoint-based blending method, where a camera image that was captured from a direction close to the current viewing direction of the user has higher influence than the colour determined via the surface normals as described above. The algorithm starts with finding the closest camera by comparing the angles between the camera directions (vector from surface to camera) and the direction to the current viewpoint (vector from surface to viewer), see Figure 5-10, 2. If the smallest angle is

below a threshold, then the image of this camera is blended over the previously computed texture, see Figure 5-10, 3. The blending factor is inversely proportional to the angle between the closest camera and viewer direction and ranges from zero to one. Smaller angles produce a higher blend-in factor and an angle of zero results in fully displaying the pixel of the closest camera.

A suitable choice for the threshold is influenced by the arrangement of capture cameras and the preference of blending behaviour. A narrow threshold causes the texture to fade-in only when the viewer is very close to a camera view, whereas a large threshold causes the texture to fade-in from a larger distance. In our setup, a threshold of 12 degrees was chosen.

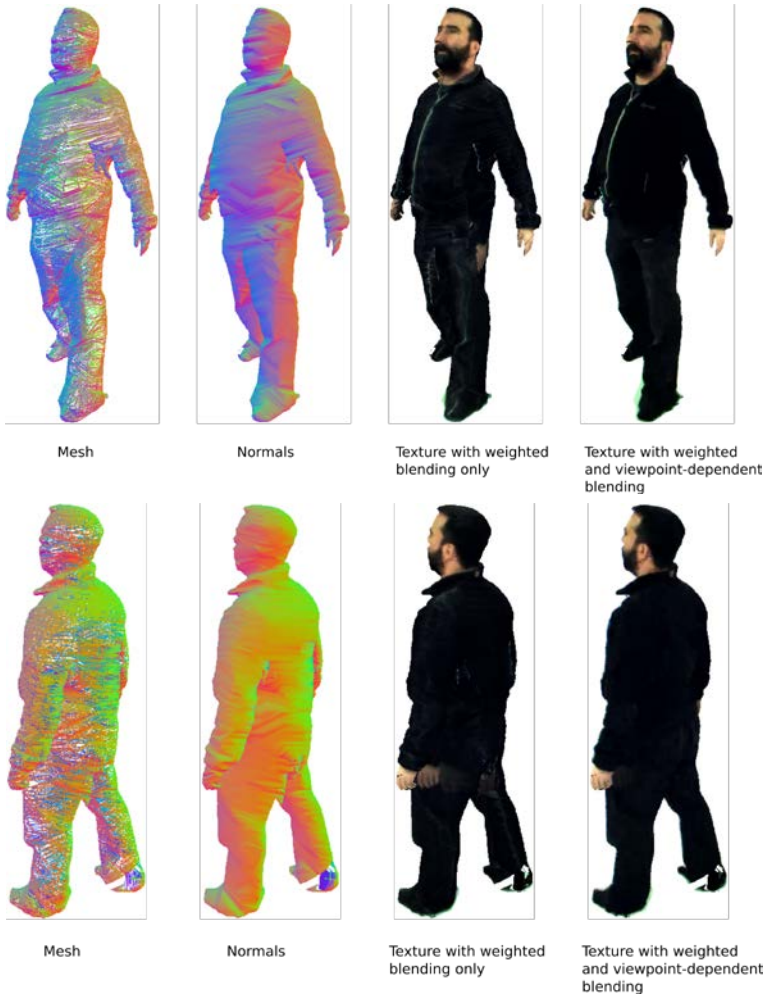


Computing the final pixel color of the texture by combining weight based blending with viewpoint dependent blending of projected camera images. 1.) Weighted blending based on the angle between surface normal N and camera vector C_i, where w_i is the weight of a camera, c_i is the projected pixel color of a camera, and n is the number of cameras. 2.) Viewpoint-dependent blending based on the angle between the viewpoint vector V and vector of the closest camera, where e_m is the smallest angle and θ is a threshold. 3.) The final color is the combination of the weighted color and the color of the closest camera to the viewpoint blended-in, where c_p is the resulting pixel color, c_w is the computed weighted color of a camera, c_m is the projected pixel color of the camera closest to the viewpoint.

Figure 5-10 Texturing Process

The result is shown in Figure 5-11. The combination of both texture mapping methods provides the best compromise of computation effort and visual quality for our system. Although, the viewpoint-based blending technique is only effective when the viewer

looks at the reconstructed mesh from near a camera view, it significantly improves the visual quality at these occasions. For example, the collar is correctly coloured in the front view and the ear is rendered with a shadow, see Figure 5-11. As our texture mapping technique does not test for visibility of surfaces to cameras, the viewpoint-based blending method has a further advantage, as it hides wrongly applied pixels to occluded areas. For example, with the simple weighted blending method based on surface normals, the pixels of the hand captured by the camera to the left of the subject are mapped onto an area of the reconstructed mesh (near the hip), see Figure 5-11. By applying the viewpoint-based texture mapping method, the image of the front camera is blended over the weights-based texture and the projected hand on the hip disappears.



Stages of 3D model rendering: incoming mesh; normal generation; texture generation via weighted blending of projected camera images; and blending of the image of the camera that is close to the user's viewpoint.

Figure 5-11 Texturing Process

5.1.7 Proof-of-Concept End-to-End Demo

The reusable code contained within the render node client has been successfully integrated with the CROSS DRIVE Virtual Environment for Mars Science Analysis (Gerndt et al., 2015) and a proof-of-concept demonstrated, where participants' 3D avatars were streamed across the Internet to multiple locations and rendered on the virtual surface of Mars. In the proof-of-concept demo the render client component of the end-to-end system presented in the thesis was integrated with a Mars Simulator (Westerteiger et al., 2011). During the live linkup 3D reconstructions of participants were generated in the

Salford University Octave and streamed to two locations simultaneously where they were rendered within the simulator. In the Octave the remote participants, represented by CGI avatars, were shown on a single screen. During the linkup the participants navigated to various locations on the simulated surface of Mars and were able to interact with one another. The architecture of the linkup is shown in Figure 5-12 and the published paper is included in Appendix B.

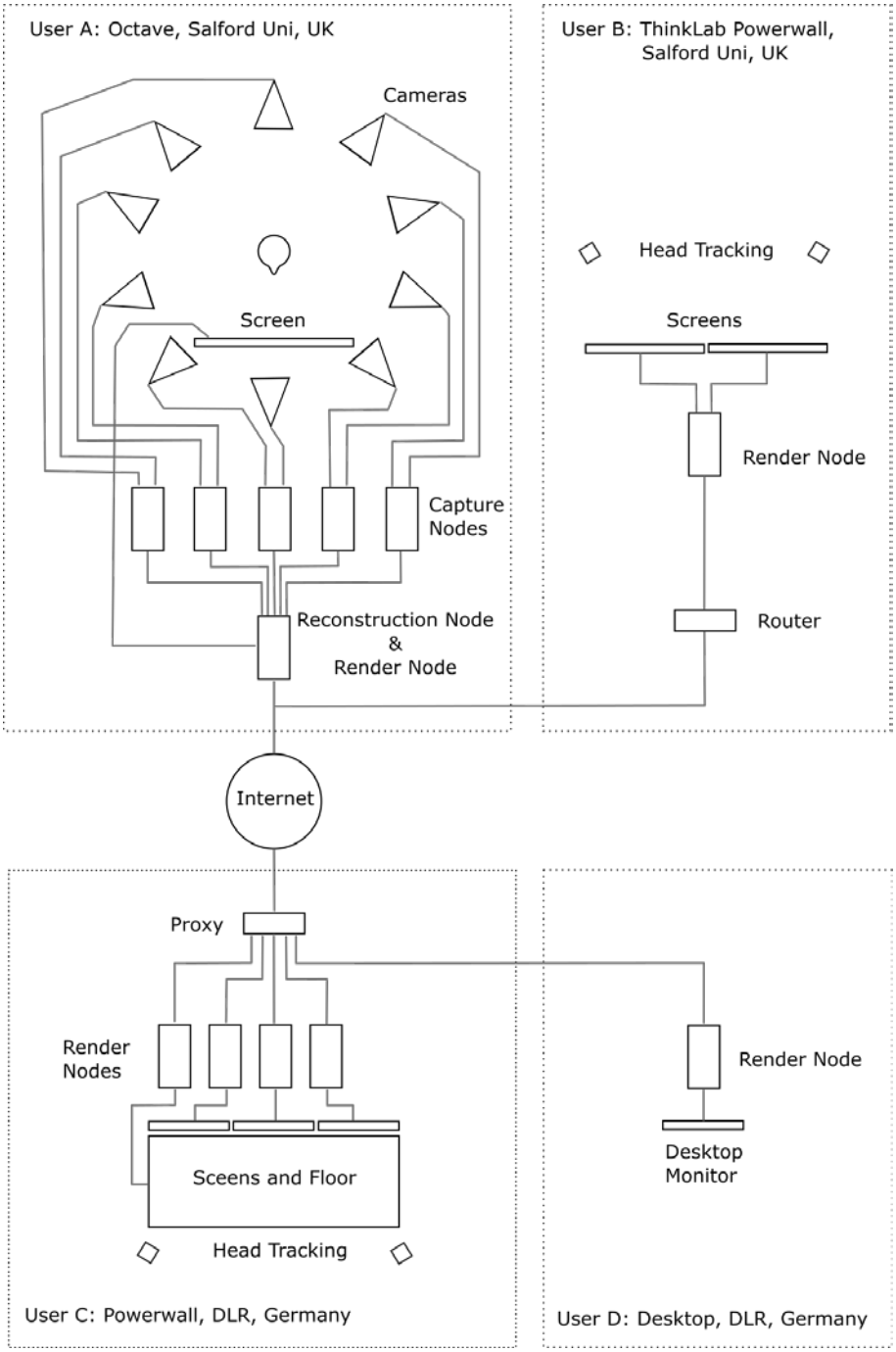


Figure 5-12 Demo Linkup Architecture

Figure 5-13 Proof-of-Concept Example One

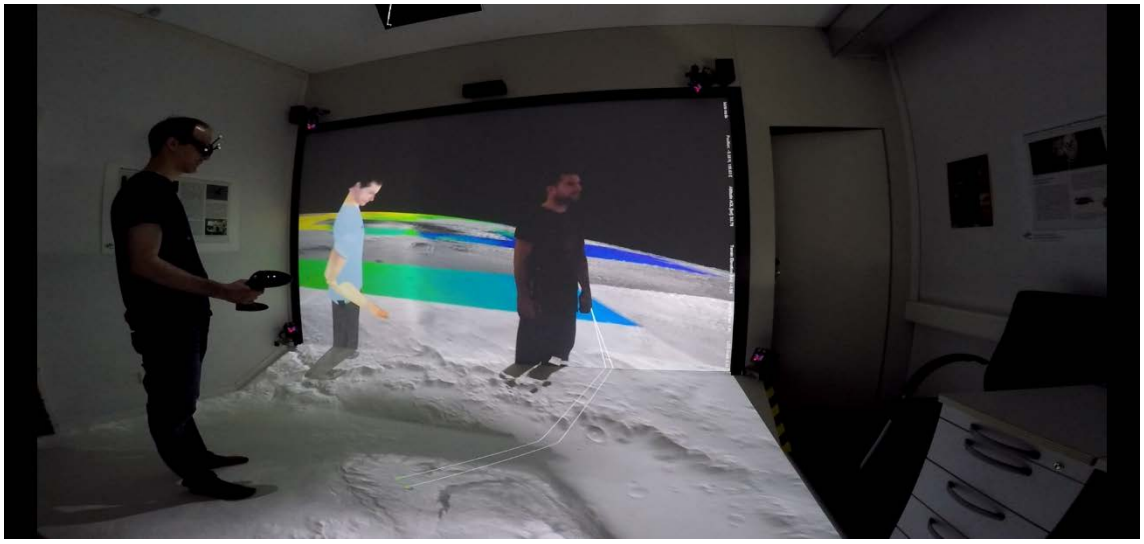


Figure 5-14 Proof-of-Concept Example Two

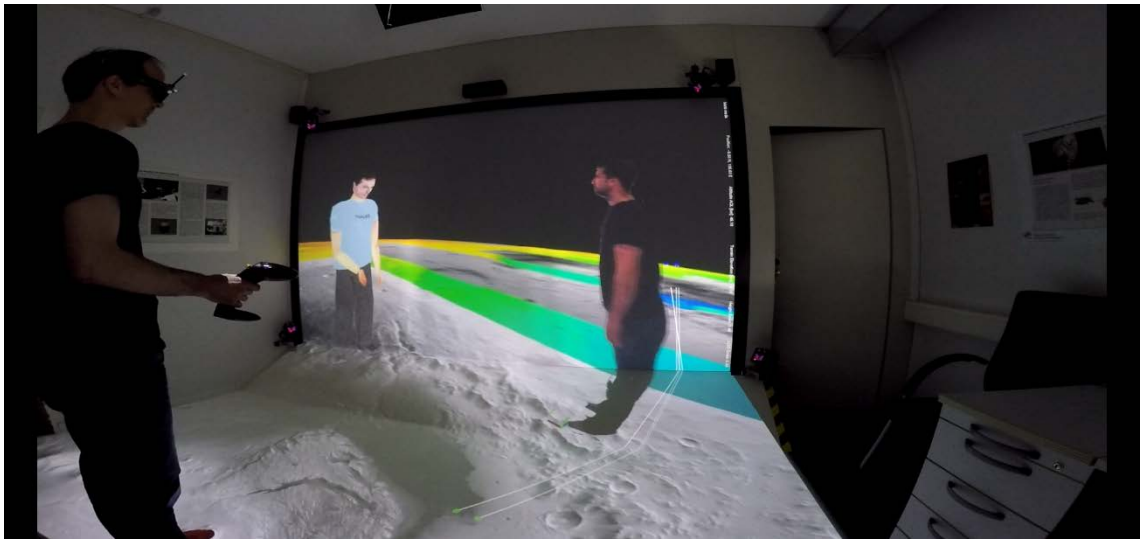


Figure 5-15 Proof-of-Concept Example Three

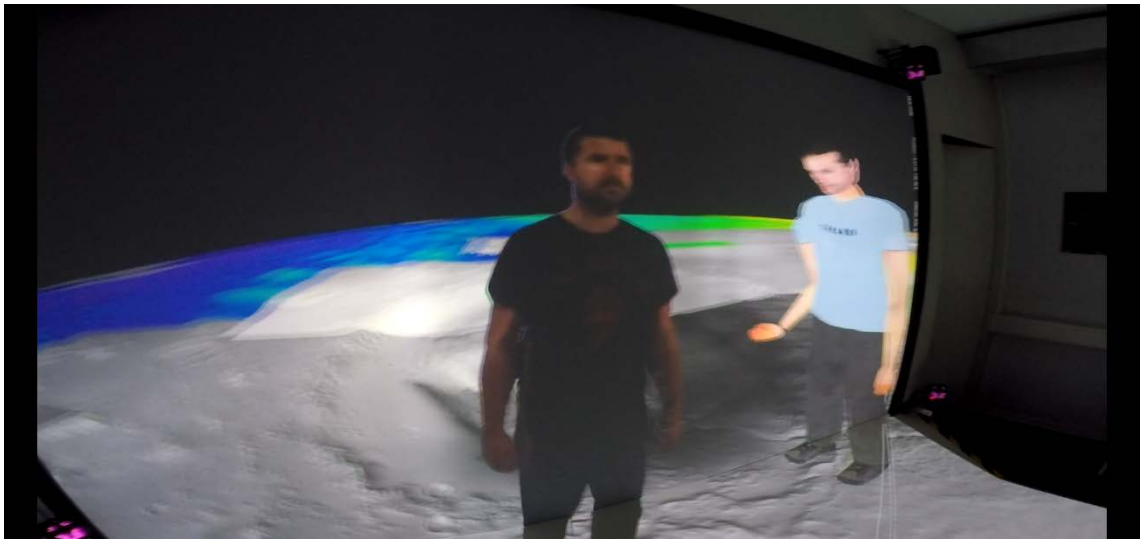


Figure 5-16 Proof-of-Concept Example Four

5.1.8 Conveying Non-Verbal Behaviour

Non-Verbal Behaviour such as gestures play an inherent role in our everyday communication, to the extent that we make use of them even when our interlocutor is not present, such as when speaking on the phone (Rimé, 1982). Gestures can be used to communicate meaningful information (semiotic), manipulate the physical world (ergotic) or even to learn through tactile exploration (epistemic) (Cadoz, 1994). Semiotic gestures have been of particular interest to the HCI community as a powerful way to communicate

with computers (McNeill, 1992; Rimé & Schiaratura, 1991). The system has demonstrated its ability to convey NVB such as pointing (Figure 5-17), waving (Figure 5-18) and interpersonal distance (Figure 5-19) as well as more subtle NVB such as eye-gaze and facial expressions as shown in which illustrates that the reconstruction is faithful enough to support conveying the seven universal emotions (Ekman & Matsumoto, 2008).

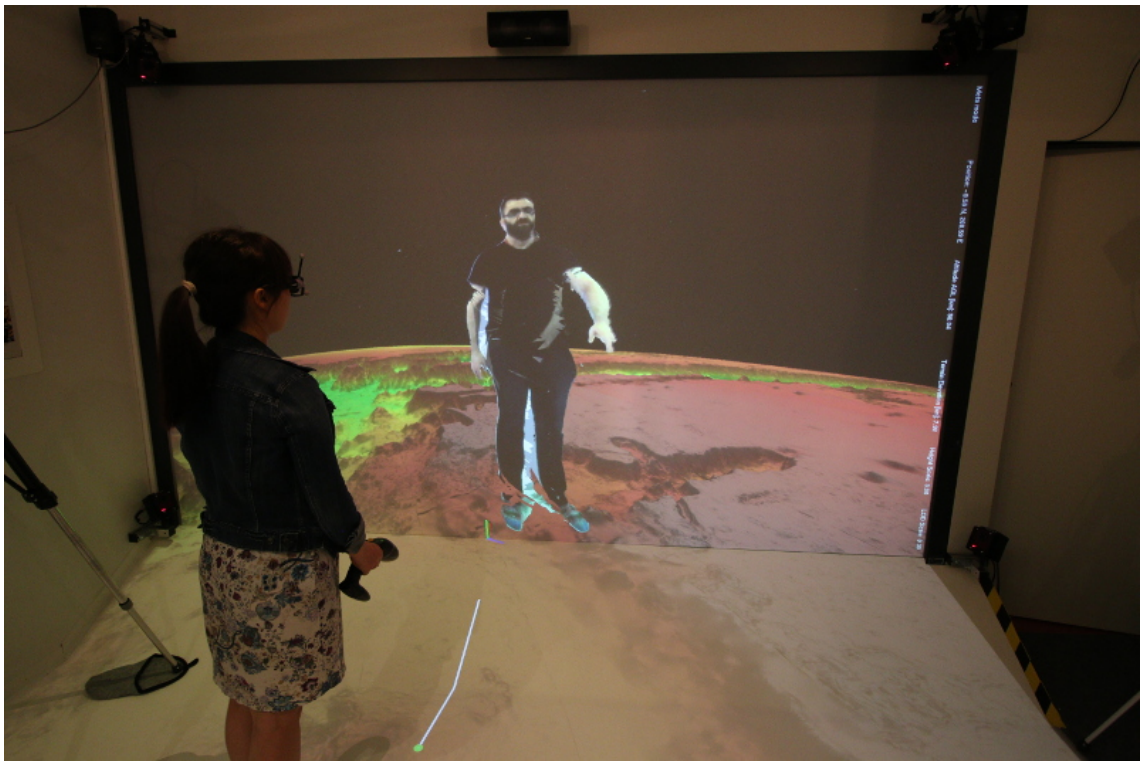


Figure 5-17 User Pointing

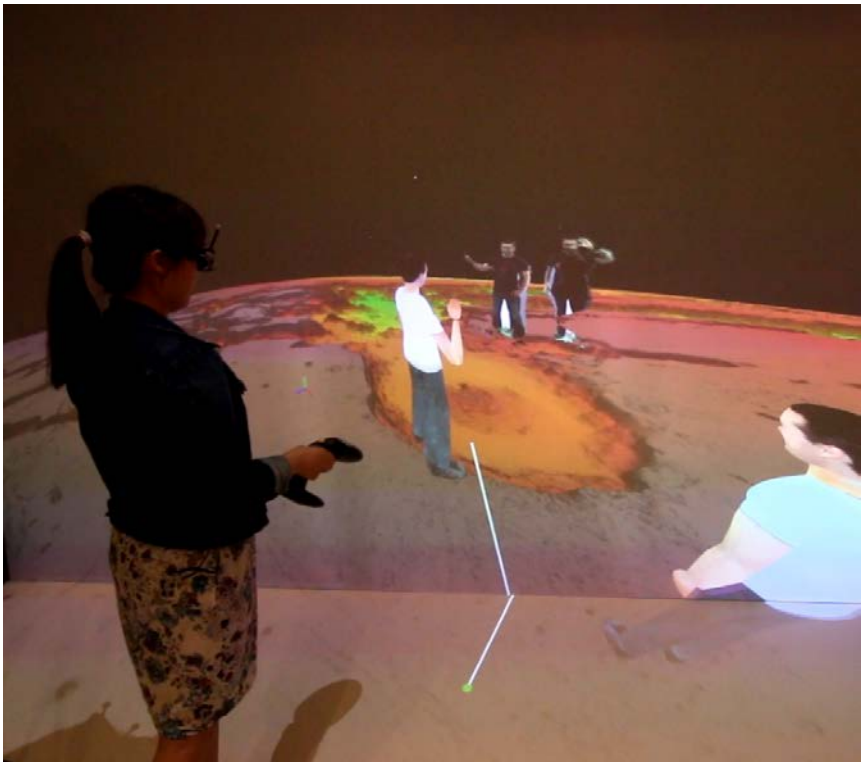


Figure 5-18 Two 3D Reconstructed Users and CGI Avatar Waving

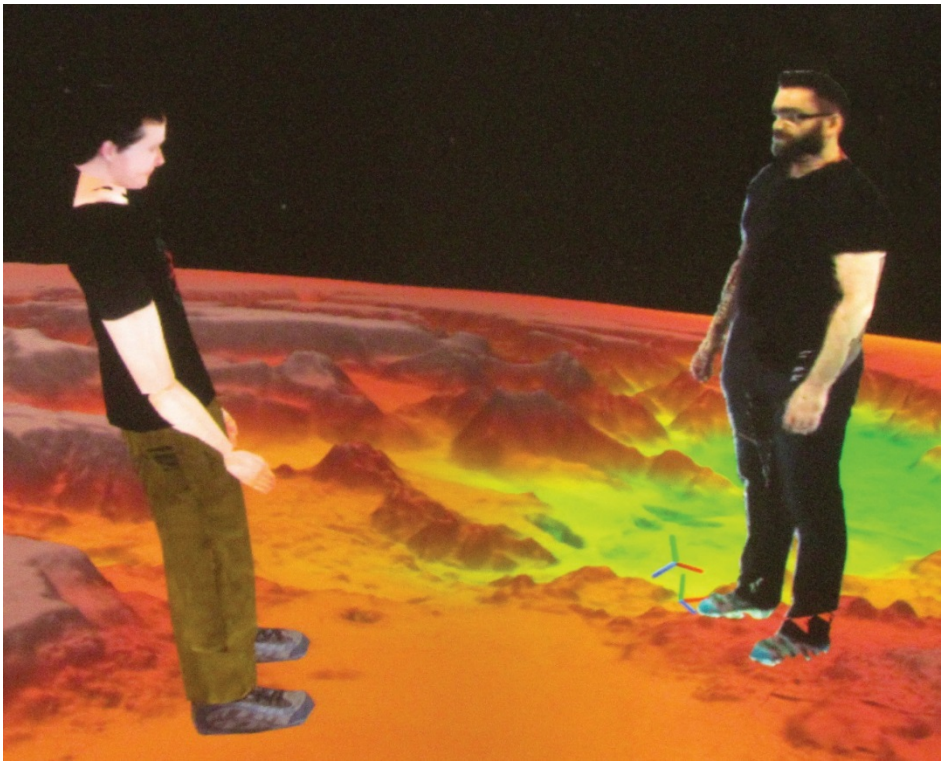


Figure 5-19 Demonstrating Interpersonal Distance Between 3D Avatar and CGI Avatar

Figure 5-20 Universal facial expressions of emotion and eye gaze. This figure shows a 3D reconstruction of the author attempting to recreate the seven universal emotions and eye gaze

It can also support the reconstruction of objects that are brought into the capture volume with the participants (Figure 5-21 through Figure 5-23).

Figure 5-21 User Holding and Writing on a Chalkboard View A

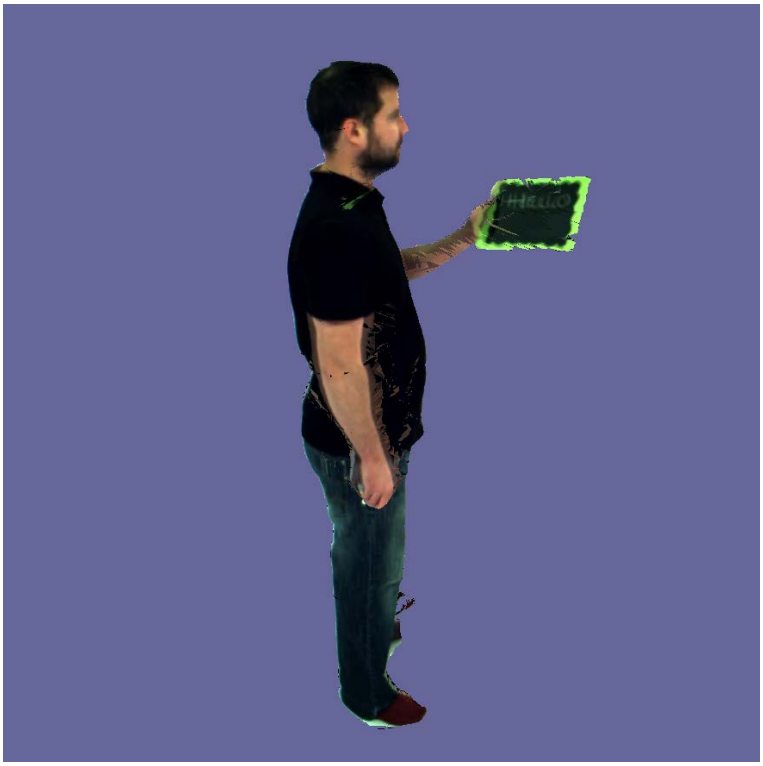


Figure 5-22 User Holding and Writing on a Chalkboard View B

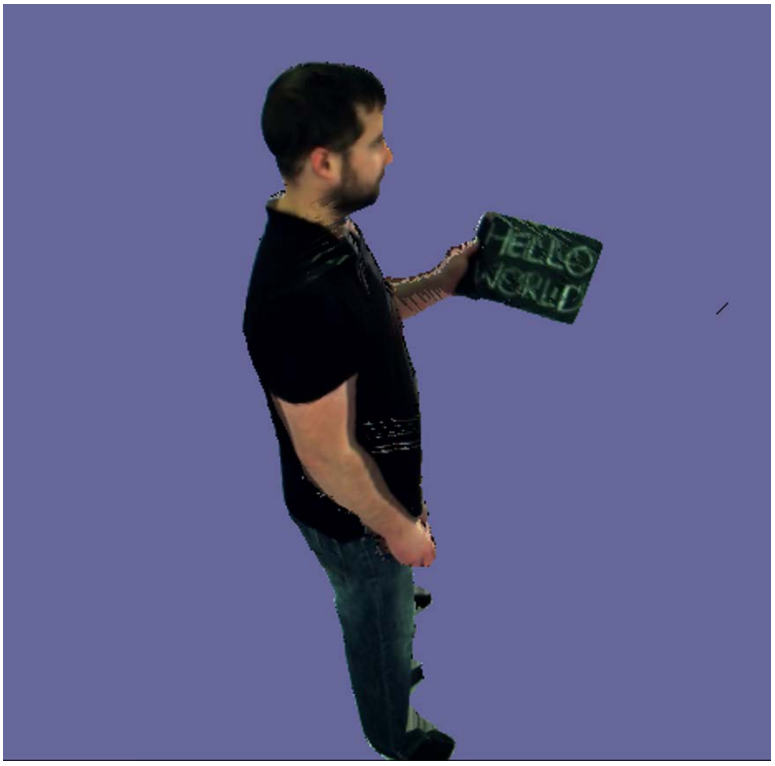


Figure 5-23 User Holding and Writing on a Chalkboard View C

Furthermore, the system has demonstrated its ability to parse pre-recorded 3D video avatars whilst integrating a live stream into a virtual environment in the context of a Virtual Reality Telepresence Exposure Therapy system. A representation of the setup is shown in Figure 5-24 and the published paper is included in Appendix C.

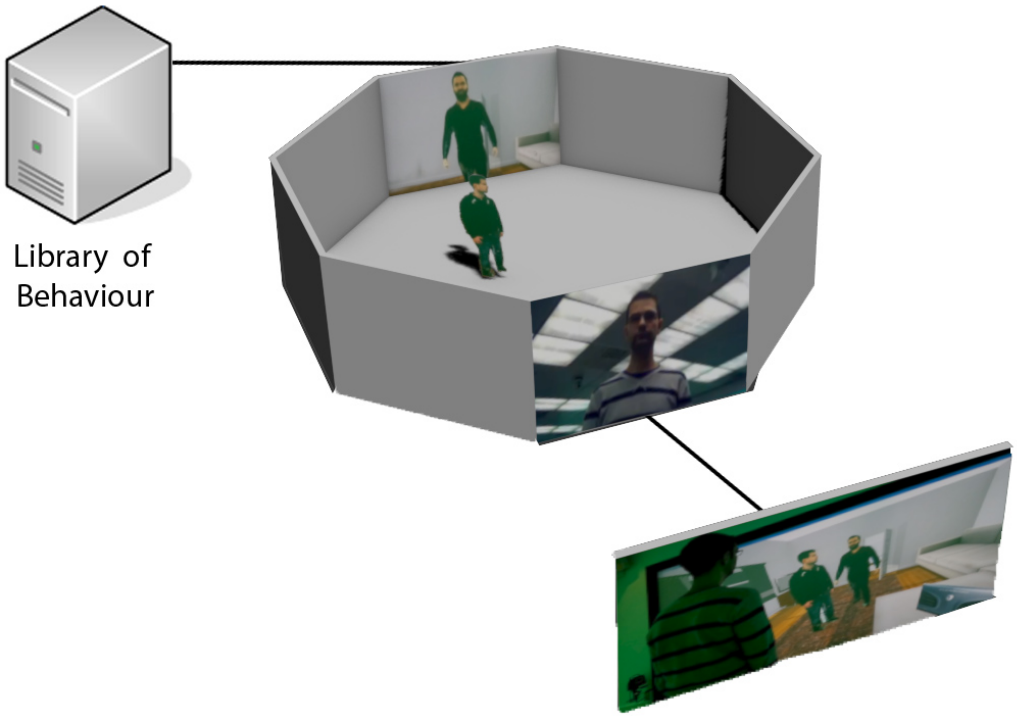


Figure 5-24 Diagram of Setup of the Asymmetric Telepresence System

And gestures conveyed are depicted in Figure 5-25 through Figure 5-27.

Figure 5-25 Client Looking at an Approaching Virtual Threat that is Prerecorded

Figure 5-26 The Therapist (Captured Live) Moves Between Client and Threat, and Tries to Redirect Client Attention (View A)

Figure 5-27 The Therapist (Captured Live) Moves Between Client and Threat, and Tries to Redirect Client Attention (View B)

The system is also being used to assist in other research at the university with the development a situated display using Chromatte for multi-view where the author would like to investigate multiple spatially distinct views of the reconstructed model. Some images taken during experimentation are shown below. Notice the camera in Figure 5-28 has been moved from the Octave wall and placed near to the user on a tripod to gain greater clarity of the users features when the remote user at angle within a threshold of that camera.

Figure 5-28 User Being Acquired in the Octave

Figure 5-29 3D Avatar of User Projected onto Chromatte Material in Remote Location (Position A)

Figure 5-30 3D Avatar of User Projected onto Chromatte Material in Remote Location (Position B)

Chapter 6

Discussion and Conclusions

This thesis concludes with a discussion about the system presented and the research that was undertaken during its development. The purpose of the system, motivation for developing, problem characteristics identified and the contributions made are revisited. This is followed by a retrospective review of how the employed methodology shaped the research. Finally, the limitations of the system and research are identified, conclusions drawn and possible future work suggested.

6.1 Discussion

This thesis presented a complete system for capturing and rendering the three-dimensional form of humans and objects that is capable of capturing in everyday environments, mixed display environments and a mixture of both. It allows the reconstructed form to be simultaneously distributed to, and rendered in, multiple locations. In addition, the form may be recorded and played back in natural time with free-viewpoint update.

The motivation for developing the system is the desire for one that can be used for research without complicated and time consuming setup, be deployed in a host of environments on commodity hardware and reproduce a wide range of non-verbal behaviour.

The system utilises components from two prototypes: Moores' (2012) distributed video system that is capable of acquiring images in real-time from multiple cameras was updated and extended. Duckworths' (2013) parallelised version of a 3D visual hull algorithm that creates models of subjects using images and corresponding silhouette information was utilised to reconstruct the 3D form.

A decision to use the prototypes was only decided following a review of the literature and similar systems including other methods that could be used to capture the 3D form.

The problem characteristics were identified as system architecture, camera calibration and background-foreground segmentation. The methodology employed for development required that progress be monitored and fed back into the process.

The following subsections discuss the contributions.

6.1.1 Spatial and Colour Calibration

6.1.1.1 Spatial

After reviewing the literature and deciding to retain the current method an understanding was gained as to why it was falling short of expectations. Proposal, implementation, testing and refinement of a new method has resulted in the system being easy to use and the calibration result consistently accurate. Documentation included in Appendix A also allows others to repeat the procedure and several users have been able to do so successfully with the author verifying this.

6.1.1.2 Colour

Two methods to correct colour were proposed and implemented. The advanced method produced best results but requires human input in its current form. Therefore, a simpler approach was integrated into the system. As well as methods to correct the colour, the updated system architecture's rendering client uses a weighted blending technique.

6.1.2 Segmentation

6.1.2.1 Visible Light Spectrum

The shortcomings of the current segmentation method employed were identified and alternative approaches discovered and tested. Finally, an approach was adopted and integrated that resulted in consistently accurate segmentation and as a result faithful reconstructions.

6.1.2.2 Infrared Spectrum

Segmenting in the infrared spectrum overcomes many of the limitations in the visible light spectrum as it is containing the segmentation problem to a very short wavelength of light that can be controlled much more effectively and the environment sterile.

6.1.3 System Architecture

The system architecture is capable of streaming avatars to multiple destinations simultaneously and recording the sessions. These sessions can then be played back in natural time. Furthermore, all calibration data so that texturing can be performed are stored with the single file reducing the requirement of keeping notes as to which calibration file goes with the recordings and thus adding to the ease of use.

The system has been successfully integrated into a Mars simulator and a live link-up demonstrated within the context of an EU funded project called CROSSDRIVE. A paper describing the link-up has also been accepted. Other researchers at the University of Salford are currently using their research and it is anticipated that further publications which the author can be attributed to will be forthcoming. The system has also been used for several undergraduate projects. It has also been successfully demonstrated on numerous occasions to visitors to the research centre which the author resides.

6.2 Limitations and Future Work

6.2.1 Spatial and Colour Calibration

6.2.1.1 Spatial

The result of the calibration process is now consistently of high quality. However, it would be of advantage if the capture system was enhanced with a feedback system to inform the calibrator when the wand spheres are in view of all cameras simultaneously. This could be achieved through audio cues where the system informs the calibrator how many cameras were observing both spheres, assisting the user with wand movements. Furthermore, if the calibration sparse bundling process were to be executed on samples of points whilst the calibrator was waving the wand, for instance at intervals of 1000

points, then it could give an indication of how accurate the result were to be so that the process could be halted once a sufficient quality was achieved. This too could be indicated by an audible cue to the calibrator.

6.2.1.2 Colour

Although the simple method of colour correction has proven effective further work could be carried out with the advanced method to make it fully autonomous. For example, the resultant silhouette from segmentation could be fed into the process to guide the process of identifying the predominant colours in the areas of interest. It may also be useful to use an optical spectrum analyser to determine the exact wavelength of the light emitted by the balls to assist with both ball choice and hue thresholding.

6.2.2 Background-Foreground Segmentation

6.2.2.1 Visible Light Spectrum

The hardware constraints imposed are not fundamental issues and it has been demonstrated that, on newer hardware, the background-foreground segmentation can in fact execute in near real-time.

Although the Ground Truthing adds value and validates the results further research could be performed to find better methods and metrics. Perhaps polygon length could be explored.

6.2.2.2 Infrared Light Spectrum

Experimenting with use of additional infrared lamps of higher quality, greater power output and the exact wavelength of the cut filter present in the cameras. An example setup is shown in Figure 6-1.

Exploring the possibility of mixing IR and visible light for the segmentation could be possible with correctly calibrated narrowband where a visible light is adjacent to an infrared camera.

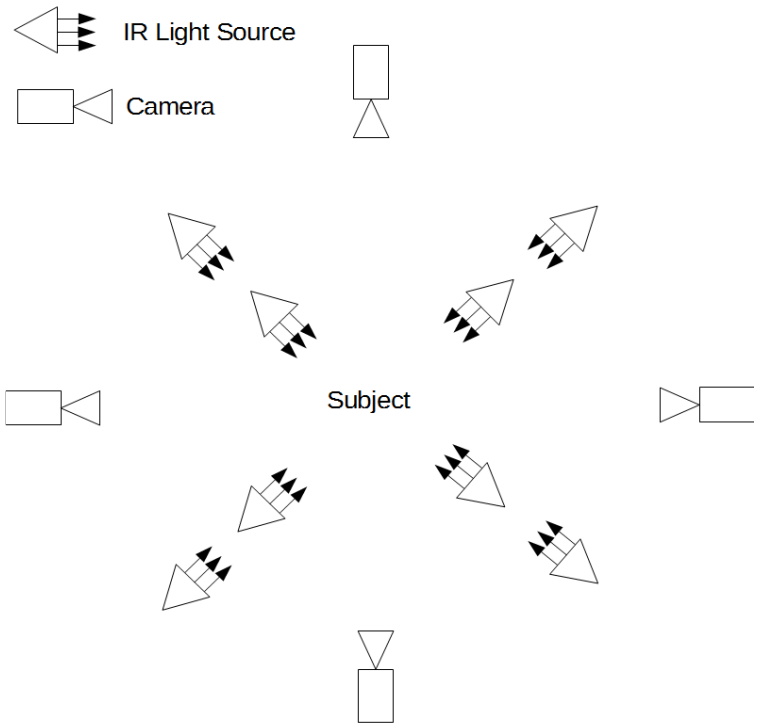


Figure 6-1 Example of Infrared Lamp and Camera Configuration

6.2.3 System Architecture

In its current deployment the system does not support micro expressions but this is due to hardware constraints and not the methods employed.

The compression of the 3D mesh is currently achieved in a relatively simple manner with each mesh being compressed in isolation from the complete stream. Further work could be done by including some form of mesh sequence encoding and compression.

The end-to-end latency of the system was measured by the author to be 1.5 second in the streaming that took place to Germany. This was measured by audio and visual cues and provides a valid indication that the delay is too great for real time systems. However, given that this was done over a TCP/IP network that one currently has no control over the latency that can be expected. The continued adoption of IPv6 and the possibility of guaranteeing bandwidth between two end-nodes provides confidence that in the future this latency can be reduced significantly.

The usage of a single PC for capture and reconstruction could be possible. Some preliminary investigation shows that if a 10Gbps capable switch and network card was present in the capture/reconstruction server and if it was equipped with multiple Graphics Cards to perform the segmentation it could reduce the requirement for multiple capture node hardware.

The links between sites are currently neither authenticated nor secure. It may be of interest for further research to investigate ways of enhancing the security without impacting on the performance.

6.3 Conclusion

This thesis has presented a complete end-to-end system capable of capturing, reconstructing, streaming and finally rendering the 3D form of people and objects. It has overcome several problem characteristics that were identified, namely system architecture, calibration and background-foreground segmentation. It enables researchers

without domain specific knowledge to investigate telepresence and NVB. It is able to perform this using commodity hardware. The thesis also presented an investigation into performing segmentation in the infrared spectrum which will provide much insight for future research.

References

7-ZIP. (2015). LZMA algorithm [Online]. Retrieved 01/09/15, from http://www.7-zip.org/

Abramov, A., Pauwels, K., Papon, J., Wörgötter, F., & Dellen, B. (2012). *Depth-supported real-time video segmentation with the Kinect.* Paper presented at the Applications of Computer Vision (WACV), 2012 IEEE Workshop on.

ACM. (2015). Retrieved 14/11/15, 2015, from http://www.acm.org

Adrian Hilton. (2015). Retrieved 14/11/15, 2015, from http://kahlan.eps.surrey.ac.uk/Personal/AdrianHilton/Welcome.html

Alexiadis, D. S., Zarpalas, D., & Daras, P. (2013). Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras. *Multimedia, IEEE Transactions on, 15*(2), 339-358.

Backlund, P., Engström, H., Hammar, C., Johannesson, M., & Lebram, M. (2007). *Sidh-a game based firefighter training simulation.* Paper presented at the Information Visualization, 2007. IV'07. 11th International Conference.

Baumgart, B. G. (1975). *A polyhedron representation for computer vision.* Paper presented at the Proceedings of the May 19-22, 1975, national computer conference and exposition.

BBC Research & Development. (2015). Retrieved 14/11/15, 2015, from http://www.bbc.co.uk/rd

BEAMING : Being in Augmented Multi-Modal Naturally-Networked Gatherings. (2013). Retrieved 14/11/15, 2015, from http://beaming-eu.org/home

Benezeth, Y., Jodoin, P.-M., Emile, B., Laurent, H., & Rosenberger, C. (2008). *Review and evaluation of commonly-implemented background subtraction algorithms.* Paper presented at the Pattern Recognition, 2008. ICPR 2008. 19th International Conference on.

blue-c. (2003). Retrieved 14/11/15, 2015, from http://blue-c.ethz.ch

blue-c-II. (2012). Retrieved 14/11/15, 2015, from http://blue-c-ii.ethz.ch

Cadoz, C. (1994). Les réalités virtuelles.

CiteSeerX. (2015). Retrieved 14/11/15, 2015, from http://citeseerx.ist.psu.edu

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2011a). *Aletheia-an advanced document layout and text ground-truthing system for production environments.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Clausner, C., Pletschacher, S., & Antonacopoulos, A. (2011b). *Scenario driven in-depth performance evaluation of document layout analysis methods.* Paper presented at the Document Analysis and Recognition (ICDAR), 2011 International Conference on.

Cockburn, A. (2008). Using both incremental and iterative development. *CrossTalk, May*.

Corporation, i. (2015). ZeroMQ. Retrieved 25/11/15, 2015, from http://zeromq.org/

Debevec, P. E., Taylor, C. J., & Malik, J. (1996). *Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach*. Paper presented at the Proceedings of the 23rd annual conference on Computer graphics and interactive techniques.

Directors of the Intel VCI (2015). Retrieved 14/11/15, 2015, from http://www.intel-vci.uni-saarland.de/en/team/

Duckworth, T. (2013). *Improving the performance of video based reconstruction and validating it within a Telepresence context.* University of Salford.

Duckworth, T., & Roberts, D. J. (2014). Parallel processing for real-time 3D reconstruction from video streams. *Journal of Real-Time Image Processing, 9*(3), 427-445.

Ekman, P., & Matsumoto, D. (2008). Facial expression analysis. *Scholarpedia, 3*(5), 4237.

Furukawa, Y., & Ponce, J. (2008). *Accurate camera calibration from multi-view stereo and bundle adjustment.* Paper presented at the Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.

Furukawa, Y., & Ponce, J. (2010). Accurate, dense, and robust multiview stereopsis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 32*(8), 1362-1376.

Gerndt, A., Gwinner, K., Fernando, T., Roberts, D., Musso, I., Basso, V., . . . Kasaba, Y. (2015). *Collaborative Virtual Environments for Mars Science Analysis and Rover Target Planning.* Paper presented at the European Planetary Science Congress

2015, held 27 September-2 October, 2015 in Nantes, France, Online at http://meetingorganizer. copernicus. org/EPSC2015, id. EPSC2015-928.

Godbehere, A. B., Matsukawa, A., & Goldberg, K. (2012). *Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation.* Paper presented at the American Control Conference (ACC), 2012.

Google. (2015). Protocol Buffers. Retrieved 25/11/2015, 2015, from https://developers.google.com/protocol-buffers/?hl=en

Google Scholar. (2015). Retrieved 14/11/15, 2015, from http://scholar.google.co.uk

Grau, O., Price, M., & Thomas, G. A. (2003). *A 3D studio production system with immersive actor feedback.* Paper presented at the ACM SIGGRAPH 2003 Sketches & Applications.

Grau, O., Pullen, T., & Thomas, G. (2004). A combined studio production system for 3-D capturing of live action and immersive actor feedback. *Circuits and Systems for Video Technology, IEEE Transactions on, 14*(3), 370-380.

Grau, O., Thomas, G. A., Hilton, A., Kilner, J., & Starck, J. (2007). *A robust free-viewpoint video system for sport scenes.* Paper presented at the 3DTV Conference, 2007.

Griesser, A., De Roeck, S., Neubeck, A., & Van Gool, L. (2005). *GPU-Based Foreground-Background Segmentation using an Extended Colinearity Criterion.* Paper presented at the Proceedings of Vision, Modeling, and Visualization (VMV) 2005.

Grimson, W. E. L., Stauffer, C., Romano, R., & Lee, L. (1998). *Using adaptive tracking to classify and monitor activities in a site.* Paper presented at the Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on.

Gross, M., Würmlin, S., Naef, M., Lamboray, E., Spagno, C., Kunz, A., . . . Lang, S. (2003). *blue-c: a spatially immersive display and 3D video portal for telepresence.* Paper presented at the ACM Transactions on Graphics (TOG).

Hansard, M., Lee, S., Choi, O., & Horaud, R. P. (2012). *Time-of-flight cameras: principles, methods and applications*: Springer Science & Business Media.

Heikkila, J., & Silvén, O. (1997). *A four-step camera calibration procedure with implicit image correction.* Paper presented at the Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on.

Henry Fuchs. (2014). Retrieved 14/11/15, 2015, from http://henryfuchs.web.unc.edu

Huang, P.-H., & Lai, S.-H. (2008). *Silhouette-based camera calibration from sparse views under circular motion.* Paper presented at the Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.

IEEE. (2015). Retrieved 14/11/15, 2015, from https://www.ieee.org

IEEE Xplore. (2015). Retrieved 14/11/15, 2015, from http://ieeexplore.ieee.org

Ilie, A., & Welch, G. (2005). *Ensuring color consistency across multiple cameras.* Paper presented at the Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on.

Incremental, U. B. (2008). Iterative Development. Dr. Alistair Cockburn, Humans and Technology. *Crosstalk May*.

Isabelle, S. K., Gilkey, R. H., Kenyon, R. V., Valentino, G., Flach, J. M., Spenny, C. H., & Anderson, T. R. (1997). *Defense applications of the CAVE (CAVE automatic virtual environment).* Paper presented at the AeroSense'97.

Joshi, N., & Jensen, H. (2004). *Color calibration for arrays of inexpensive image sensors.* Master's thesis, Stanford University Department of Computer Science.

KaewTraKulPong, P., & Bowden, R. (2002). An improved adaptive background mixture model for real-time tracking with shadow detection *Video-based surveillance systems* (pp. 135-144): Springer.

Larman, C., & Basili, V. R. (2003). Iterative and incremental development: A brief history. *Computer*(6), 47-56.

Laurentini, A. (1994). The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 16*(2), 150-162.

Lee, S.-Y., Kim, I.-J., Ahn, S. C., Ko, H., Lim, M.-T., & Kim, H.-G. (2004). *Real time 3D avatar for interactive mixed reality.* Paper presented at the Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry.

Li, L., Huang, W., Gu, I. Y., & Tian, Q. (2003). *Foreground object detection from videos containing complex background.* Paper presented at the Proceedings of the eleventh ACM international conference on Multimedia.

Maimone, A., & Fuchs, H. (2011a). *Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras*. Paper presented at the Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on.

Maimone, A., & Fuchs, H. (2011b). *A First Look at a Telepresence System with Room-Sized Real-Time 3D Capture and Large Tracked Display*. Paper presented at the International Conference on Artificial Reality and Telexistence (ICAT), Osaka (Japan).

Maimone, A., & Fuchs, H. (2012a). *Real-Time Volumetric 3D Capture of Room-Sized Scenes for Telepresence*. Paper presented at the Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Zurich (Switzerland).

Maimone, A., & Fuchs, H. (2012b). *Reducing interference between multiple structured light depth sensors using motion.* Paper presented at the Virtual Reality Short Papers and Posters (VRW), 2012 IEEE.

Massie, T. H., & Salisbury, J. K. (1994). *The phantom haptic interface: A device for probing virtual objects.* Paper presented at the Proceedings of the ASME winter annual meeting, symposium on haptic interfaces for virtual environment and teleoperator systems.

Matsuyama, T., Wu, X., Takai, T., & Wada, T. (2004). Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video. *Circuits and Systems for Video Technology, IEEE Transactions on, 14*(3), 357-369.

Matusik, W., Buehler, C., & McMillan, L. (2001). *Polyhedral visual hulls for real-time rendering*: Springer.

Matusik, W., Buehler, C., Raskar, R., Gortler, S. J., & McMillan, L. (2000). *Image-based visual hulls.* Paper presented at the Proceedings of the 27th annual conference on Computer graphics and interactive techniques.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*: University of Chicago press.

Mester, R., Aach, T., & Dümbgen, L. (2001). Illumination-invariant change detection using a statistical colinearity criterion *Pattern recognition* (pp. 170-177): Springer.

Mitchelson, J., & Hilton, A. (2003). Wand-based multiple camera studio calibration. *Center Vision, Speech and Signal Process*.

Moore, C. (2012). *Distribution and Processing of Video for Real-time 3D Telepresence*. University of Salford.

O'Hare, J. Octave - technical information, University of Salford. Retrieved 01/09/15, 2015, from http://www.salford.ac.uk/computing-science-engineering/facilities/octave-technical-information

Office of the Future. (2009). Retrieved 14/11/15, 2015, from http://www.cs.unc.edu/Research/stc

Petit, B., Lesage, J.-D., Menier, C., Allard, J., Franco, J.-S., Raffin, B., . . . Faure, F. (2009). Multicamera real-time 3d modeling for telepresence and remote collaboration. *International journal of digital multimedia broadcasting, 2010*.

Pollefeys, M., Sinha, S. N., Guan, L., & Franco, J.-S. (2009). Multi-view calibration, synchronization, and dynamic scene reconstruction. *Multi-Camera Networks: Principles and Applications*, 29-75.

Porikli, F. (2003). *Inter-camera color calibration by correlation model function.* Paper presented at the Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on.

Ramanathan, P., Steinbach, E. G., & Girod, B. (2000). *Silhouette-Based Multiple-View Camera Calibration.* Paper presented at the VMV.

Reflecmedia. (2015). Chromatte. Retrieved 15/12/15, 2015, from http://www.reflecmedia.com/broadcast/products/chromatte/index.htm

ResearchGate. (2015). Retrieved 14/11/15, 2015, from http://www.researchgate.net

Rimé, B. (1982). The elimination of visible behaviour from social interactions: Effects on verbal, nonverbal and interpersonal variables. *European journal of social psychology, 12*(2), 113-129.

Rimé, B., & Schiaratura, L. (1991). Gesture and speech.

Roberts, D. J., Fairchild, A. J., Campion, S. P., O'Hare, J., Moore, C. M., Aspin, R., . . . Tecchia, F. (2015). withyou—An Experimental End-to-End Telepresence System Using Video-Based Reconstruction. *Selected Topics in Signal Processing, IEEE Journal of, 9*(3), 562-574.

Roberts, D. J., Rae, J., Duckworth, T. W., Moore, C. M., & Aspin, R. (2013). Estimating the gaze of a virtuality human. *Visualization and Computer Graphics, IEEE Transactions on, 19*(4), 681-690.

Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern recognition, 33*(2), 225-236.

Schultz, C. (2006). Digital Keying Methods. University of Bremen Center for Computing Technologies.

ScienceDirect. (2015). Retrieved 14/11/15, 2015, from http://ieeexplore.ieee.org

Shen, R., Cheng, I., & Basu, A. (2008). *Multi-Camera Calibration Using a Globe.* Paper presented at the The 8th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras-OMNIVIS.

Shu, B., Qiu, X., & Wang, Z. (2008). *Hardware-based camera calibration and 3D modelling under circular motion.* Paper presented at the Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on.

Shujun, Z., Cong, W., Xuqiang, S., & Wei, W. (2009). *DreamWorld: CUDA-accelerated real-time 3D modeling system.* Paper presented at the Virtual Environments, Human-Computer Interfaces and Measurements Systems, 2009. VECIMS'09. IEEE International Conference on.

Sinha, S. N., & Pollefeys, M. (2004). *Synchronization and calibration of camera networks from silhouettes.* Paper presented at the Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on.

Sobral, A. (2013). *BGSLibrary: An opencv c++ background subtraction library.* Paper presented at the IX Workshop de Visao Computacional (WVC'2013), Rio de Janeiro, Brazil.

Starck, J., Maki, A., Nobuhara, S., Hilton, A., & Matsuyama, T. (2009). The multiple-camera 3-D production studio. *Circuits and Systems for Video Technology, IEEE Transactions on, 19*(6), 856-869.

Stauffer, C., & Grimson, W. E. L. (1999). *Adaptive background mixture models for real-time tracking.* Paper presented at the Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.

Stauffer, C., & Grimson, W. E. L. (2000). Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22*(8), 747-757.

Stockman, G. C., Chen, S. W., Hu, G., & Shrikhande, N. (1988). Sensing and recognition of rigid objects using structured light. *Control Systems Magazine, IEEE, 8*(3), 14-22. doi: 10.1109/37.472

Tong, J., Zhou, J., Liu, L., Pan, Z., & Yan, H. (2012). Scanning 3d full human bodies using kinects. *Visualization and Computer Graphics, IEEE Transactions on, 18*(4), 643-650.

Tsai, R. Y. (1987). A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *Robotics and Automation, IEEE Journal of, 3*(4), 323-344.

van den Bergh, F., & Lalioti, V. (1999). Software chroma keying in an immersive virtual environment. *South African Computer Journal, 24*(155-162), 50.

Westerteiger, R., Gerndt, A., & Hamann, B. (2011). *Spherical Terrain Rendering using the hierarchical HEALPix grid*. http://dx.doi.org/10.4230/OASIcs.VLUDS.2011.13

Will, P. M., & Pennington, K. S. (1971). *Grid coding: a preprocessing technique for robot and machine vision*. Paper presented at the Proceedings of the 2nd international joint conference on Artificial intelligence, London, England.

Yasuda, K., Naemura, T., & Harashima, H. (2003). *Thermo-key: Human Region Segmentation from Video Using Thermal Information.* Paper presented at the ACM SIGGRAPH.

Yasuda, K. N., T. ; Harashima, H. . (2004). Thermo-key: human region segmentation from video. *Computer Graphics and Applications, IEEE, 24*(1), 26-30. doi: 10.1109/MCG.2004.1255805

Zhang, H., Wong, K.-Y. K., & Zhang, G. (2007). Camera calibration from images of spheres. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29*(3), 499-502.

Zhang, Z. (2000). A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22*(11), 1330-1334.

Zhang, Z. (2004). Camera calibration with one-dimensional objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 26*(7), 892-899.

Zivkovic, Z. (2004). *Improved adaptive Gaussian mixture model for background subtraction.* Paper presented at the Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on.

Zivkovic, Z., & van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, *27*(7), 773-780.

Appendix A

3D Reconstruction System User Guide

Revision: 1

Date: 13/06/2016

Contents

About 3
Key to notations 4
Spatial Calibration of Cameras 5
Acquiring sphere coordinates 5
Capture Node PC Preparation 5
Ready the Octave 5
Acquire the points 5
Generating Calibration File 7
Running 3D Reconstruction and Rendering 9
Configuration of octave 9
For optimal results in the Octave research facility turn on all ceiling lights and set all wall projectors to white. It is best to leave the projectors which illuminate the floor OFF. 9
Synchronous video acquisition and background-foreground segmentation 9
3D Reconstruction 10
Initialising the software 10
Performing 3D Reconstruction 11
3D Rendering 12
Live 12
Parsing from disk 13

About

The 3D Reconstruction System comprises of multiple network connected components. It is capable of capturing, reconstructing and rendering people and objects.

Key to notations

The following notations are used in the following document:

Note: - notes appear like this

Important: - important notes appear like this

RunThisExecutableOrBatchProcessorFile - indicates a file that should be executed

'Button click' - denotes the name of a button that can be clicked

WindowName - identifies the name of a window or dialogue

CHANGEME - indicates a value in a configuration XML file that may be changed

Spatial Calibration of Cameras

The process generates an XML file containing the calibration data. The calibration data contains the pose of all cameras within the capture volume. It is achieved by waving a wand with two spheres separated by known distance throughout the capture volume then evaluating the spheres coordinates. The accuracy of spatial calibration is directly linked to the quality of the models generated by the reconstruction process.

Acquiring sphere coordinates

Procedure is possible with only one person but having two present is advised.

Capture Node PC Preparation

Turn on and log in as 'Capture' on PCs SU001 – SU006

On SU002 open Windows Explorer and navigate to:
C:\Users\capture\Dropbox\SilhouetteSender\Current

Execute the batch file named: ***StartAllCamServersLoop.bat***

Wait momentarily whilst the script starts the capture software on the five capture nodes.

Once all nodes are ready, still on SU002, execute ***StartCentralServerCalibration.bat***

Ready the Octave

Turn the centre two rows of lights in the Octave 'off'.

Note: for best results you can turn all Octave lights 'off' and leave only the light in the server room 'on'. However, be mindful of low visibility in this setting.

Illuminate the wand by twisting the spheres clockwise.

If two people are present one can stand in the centre of the Octave with the wand spheres in view of all cameras.

If one person is present then place the wand on top of a foam square in the centre of the Octave.

Acquire the points

On SU002 click the ***'Start'*** button on the *NetworkGUI* window (Figure 1 NetworkGUI Window) that should still have focus:

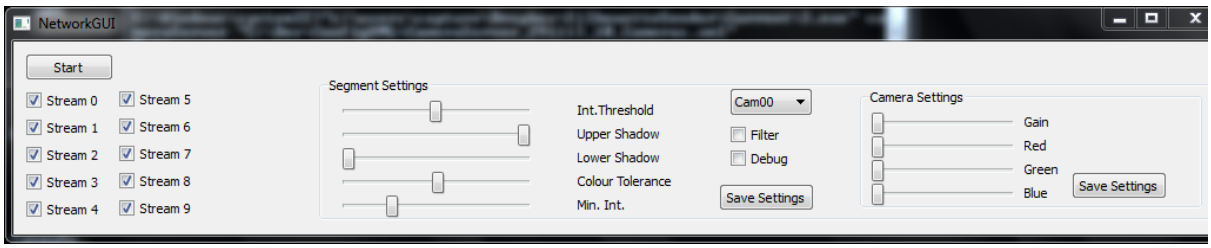


Figure 1 NetworkGUI Window

This will start a continuous and synchronised sphere coordinate locator for each camera on all the capture nodes. Each node will display two windows one for each camera (Figure 2 Windows Displaying Output of Each Connected Camera).

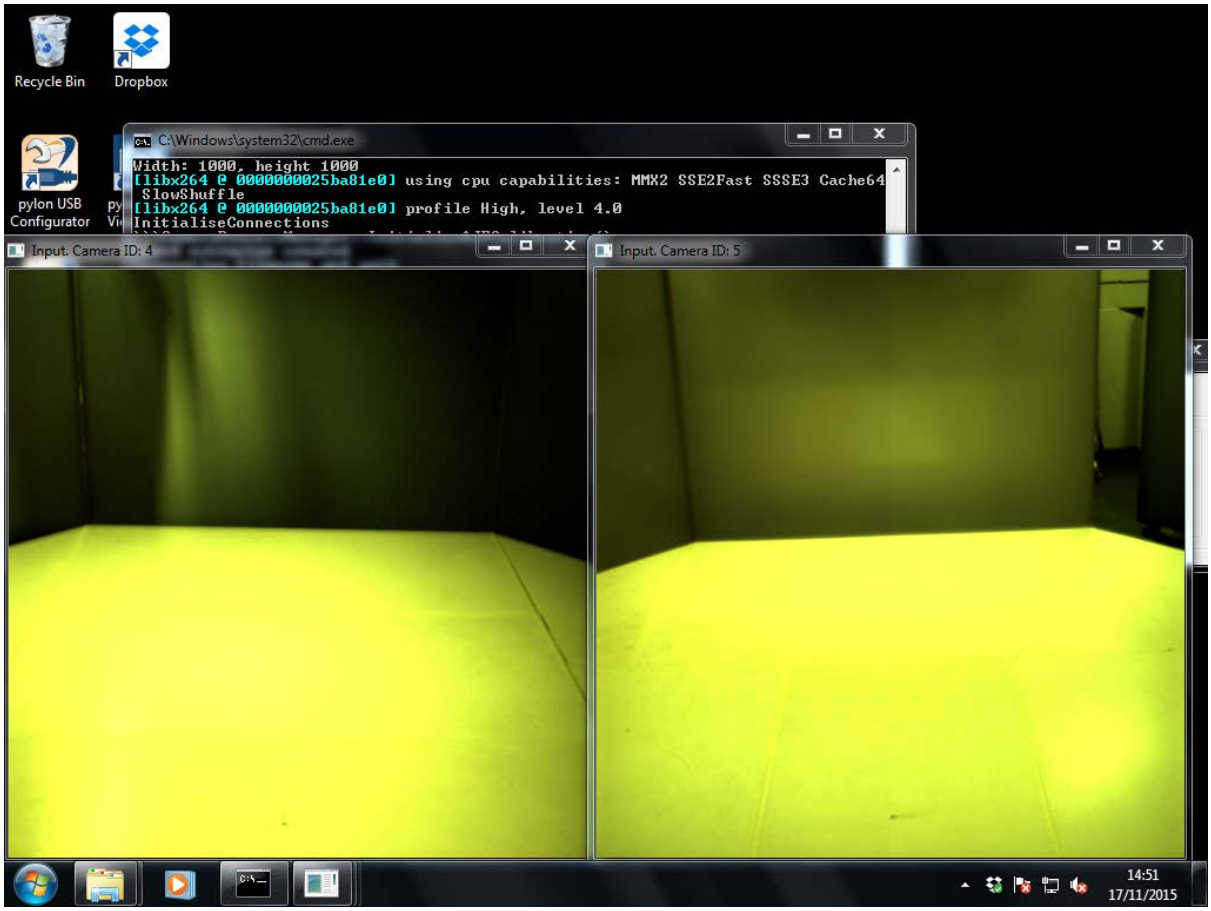


Figure 2 Windows Displaying Output of Each Connected Camera

If one person is present check that the spheres on the wand situated on the foam block in the Octave are being detected by observing each camera window. Walk to the Octave and begin waving the wand.

If two people are present inform the person present in the Octave to begin waving the wand and observe that the sphere coordinates are being located successfully.

Wave the wand throughout the capture volume. Move the wand at a steady pace and be mindful that the wand should be visible to all cameras simultaneously. It is useful to imagine the intersecting cones of the cameras views. Lying down will allow the wand to be waved in the lower segment of the volume. Allow around five minutes of waving.

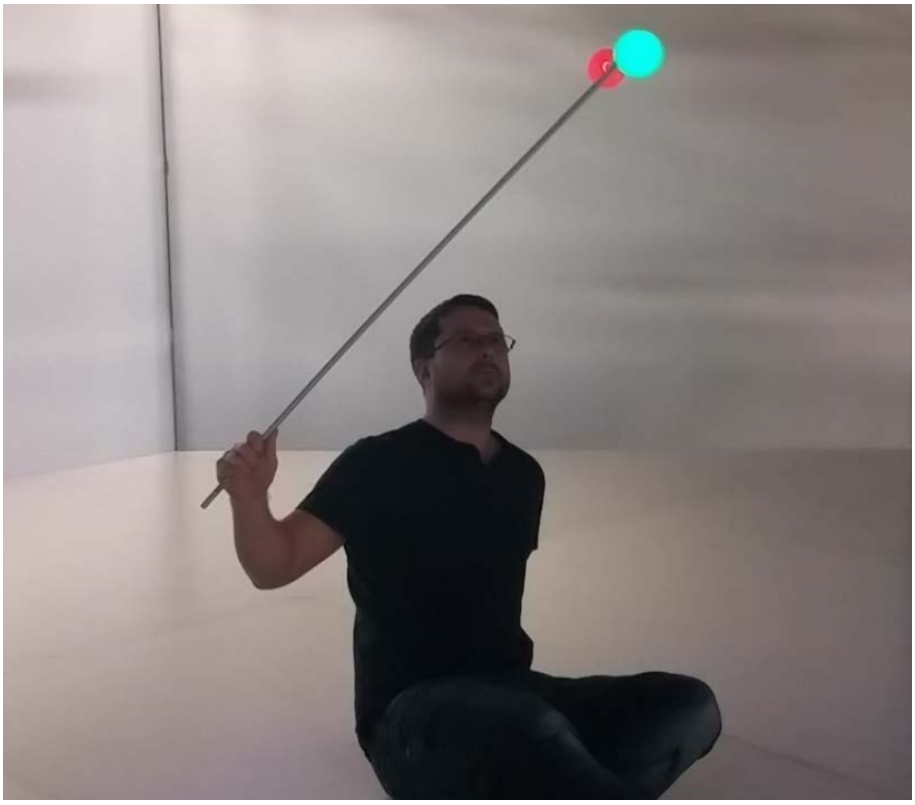


Figure 3 Waving of the Wand

If one person is present place the wand back on the foam block and return to the capture node PCs

If two present signal to the other that calibration complete but continue to wave the wand

Remaining on SU002, click on the ***'red cross'*** to close the *NetworkGUI* window. This will disconnect and close all instances of the capture software on the other PCs. The software will restart on each PC ready for the next connection.

The process of acquiring the coordinates is now complete. Ten point files points00.txt – points09.txt will be present in:

C:\Users\capture\Dropbox\CalibrationPoints

Generating Calibration File

On a PC running a 64 bit Windows Operating System and at least 8GB of RAM copy the point files into a new folder in Dropbox\Calibration labelled with the date of point acquisition (DDMMYY). Copy CalibrateWand.exe and intrinsics.txt from the root of Dropbox\Calibration into the folder with the point files and open a command window in the directory of the folder (hold down shift and right click. Then select Open Command Window here).

Run the following:

CalibrateWand.exe –files –maxframes x

Where:

-files: ensures the tool will output the result as a text file.

-maxframes x: allows the user to adjust the maximum number of frames processed. This defaults to 1000 and can be anything up to the maximum number of points acquired divided by two (two sets of points per frame) and this figure is at the top of any of the point files. Important: remember to divide the number of points in the points file by two to produce the absolute maximum number of frames.

And other parameters of noteworthiness are:

-wandlength y: this is the distance (in metres) separating the centre of the two spheres on the wand. The default is 0.485 which is the current value so the parameter can be omitted.

Once the calibration process is complete a preview of the calibration file (in XML format) will be outputted to the command window.

Two new files will be written to the same directory:

calibrate_wand_stats.txt – which contains statistics

calibration.xml – contains the calibration data for the cameras current configuration to be used in the 3D reconstruction software

Note: the CalibrateWand software will always produce a result that is useable, however, it is advised to check the quality of the result. To do this check the output in calibrate_wand_stats.txt. If it looks like:

-1 -1 -1

and the command line output shows 'ERROR: block inversion failed (bad constraint?)' similar to Figure 4 Calibration Command Line Output:

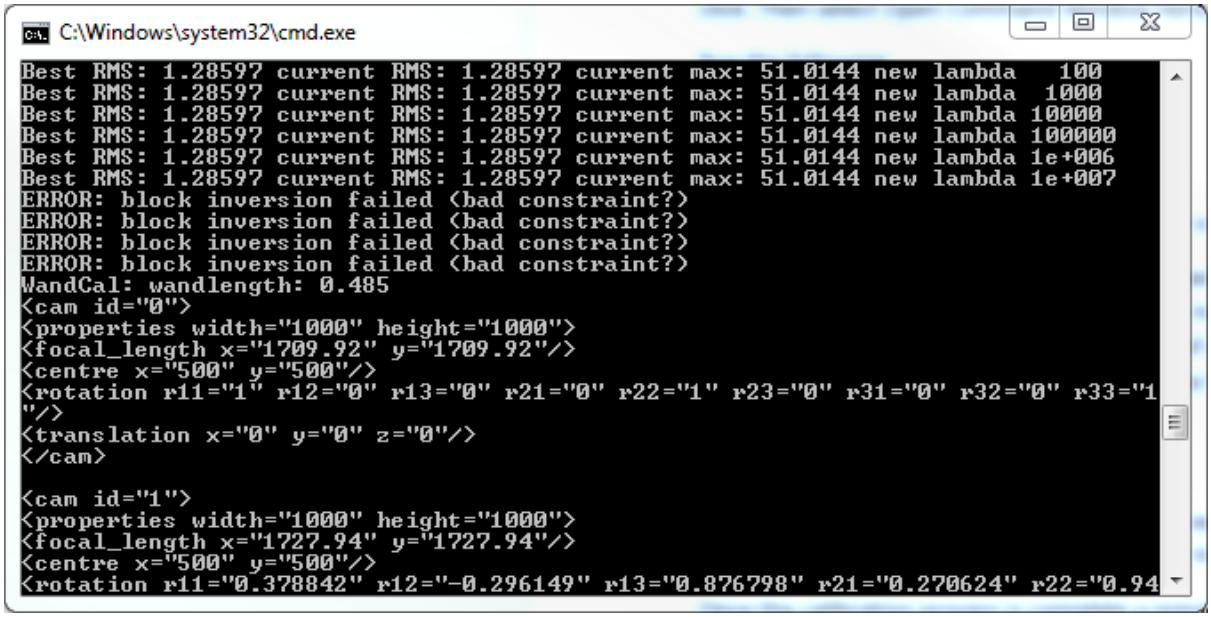


Figure 4 Calibration Command Line Output

then there was a problem with the calibration and it may have to be repeated.

However, before repeating try running CalibrateWand using more or less frames via the -maxframes argument and check the result.

Running 3D Reconstruction and Rendering

The 3D Reconstruction and Rendering software comprises of three network connected components: Synchronous video acquisition, 3D reconstruction and 3D Rendering (Figure 5 System Architecture). It is recommended to run the 3D Reconstruction component on a PC with at least 8GB of RAM and a CPU(s) with a minimum of 8 cores (4 cores on a processor with hyper threading). All software components require a 64 bit Windows Operating System.

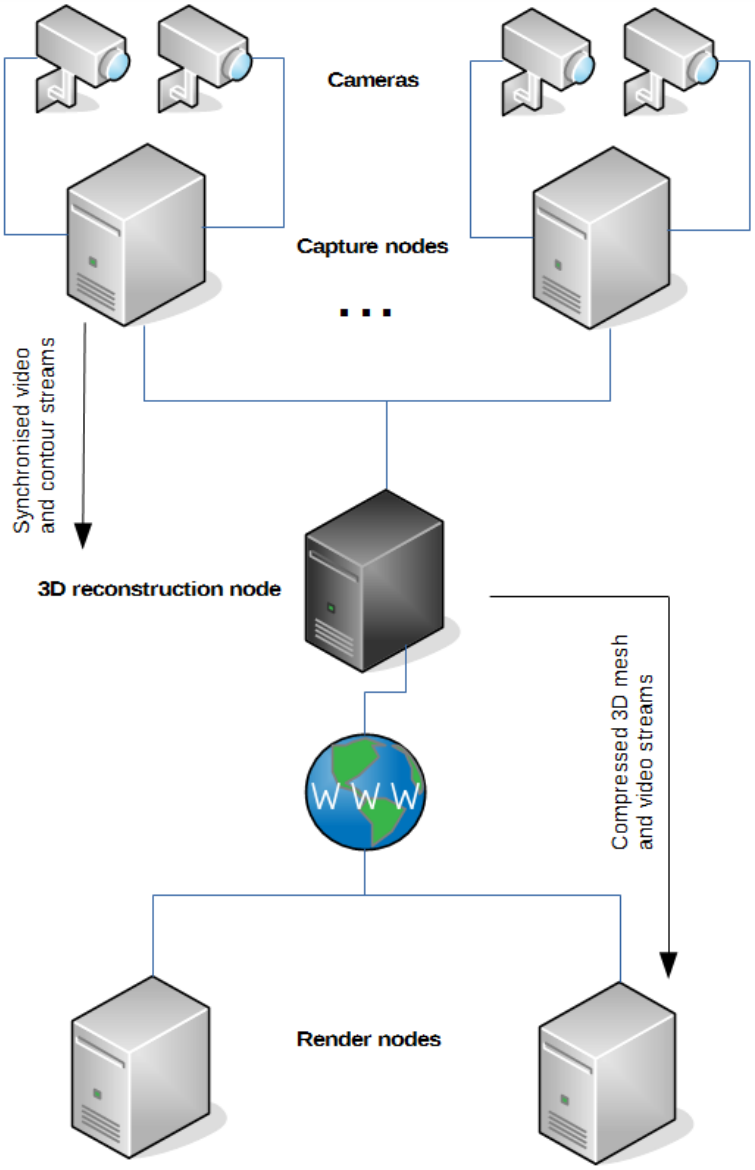


Figure 5 System Architecture

Configuration of octave

For optimal results in the Octave research facility turn on all ceiling lights and set all wall projectors to white. It is best to leave the projectors which illuminate the floor OFF.

Synchronous video acquisition and background-foreground segmentation

Important: if the capture node software is already running on the nodes then you may skip these steps.

This process runs on the five capture nodes: SU001, SU003, SU004, SU005 and SU006.

Turn on and log in as 'Capture' on PCs SU001 – SU006

On SU002 open Windows Explorer and navigate to:
C:\Users\capture\Dropbox\SilhouetteSender\Current

Execute the batch file named: ***StartAllCamServersLoop.bat***

Wait momentarily whilst the script starts the capture software on the five capture nodes.

3D Reconstruction

It is recommended to run the 3D Reconstruction software on a high performance PC.

Initialising the software

The software can run in several modes:

Standalone

Due to legacy components it is possible to view the rendered 3D reconstruction using a deprecated viewer.

To use the deprecated method of rendering execute *3DRecon.exe*

In conjunction with 3D Rendering Client(s)

In this mode the 3D Reconstruction software streams the 3D mesh and video data to connected 3D Rendering Clients.

It can be invoked by executing the batch file named: '***run_with_geom_stream – no save.bat'*** and then once the 3D reconstruction software has loaded clicking the **'Start Geom Stream'** button in the *Camera Control* dialogue (Figure 6 Camera Control Window window that takes focus.

Important: do not forget to click the ***'Start Geom Stream'*** button.

In conjunction with 3D Rendering Client(s) + saving 3D stream to disk

As well as being able to stream to the 3D Rendering Client(s), in this mode, the 3D Reconstruction software writes the 3D mesh and video data to a local disk.

To start the software in this mode execute: '***run_with_geom_stream.bat'***

It will write to a file to the directory configured in the batch processor file, for example:

3DRecon.exe -geom_stream tcp://127.0.0.1:51215 -saveprotobuf J:\ProtobufMsgs

The filename will be in the following format: geomstream_YYYY_MM_DD_HH_MM_ss.pb. For example:

geomstream_2015_10_05_10_13_46.pb

Note: it is recommended that the disk used is the one with the fastest write speed available in your system.

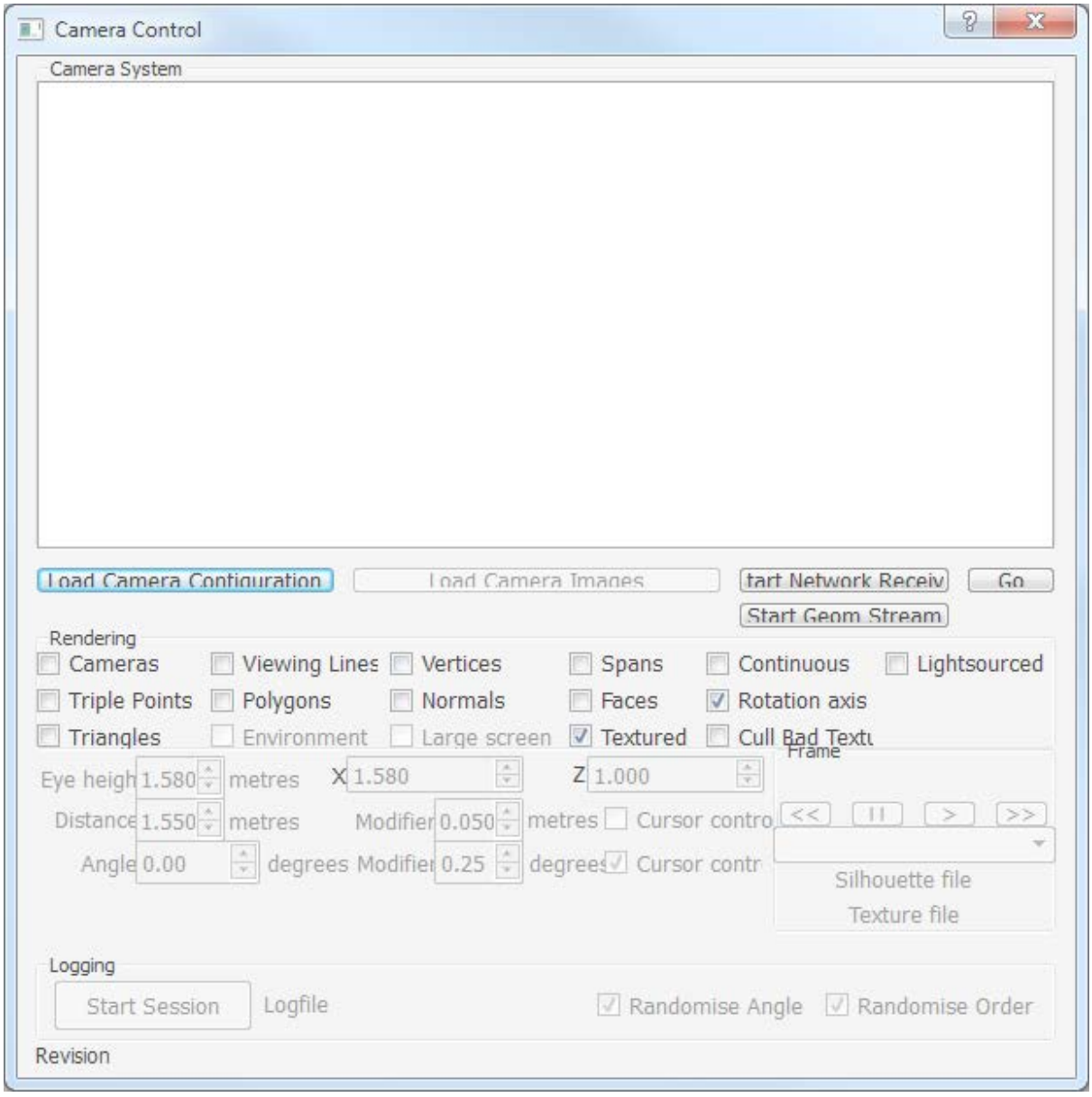


Figure 6 Camera Control Window

Performing 3D Reconstruction

Once the software is initialised in the desired mode the reconstruction process can be started.

Important: if intent on streaming to the 3D Renderer Client ensure that all instances required are running as detailed in 3D Rendering before completing either of the following:

Starting live streaming

First, ensure that all instances of the capture node software are started. Then in the Camera Control dialogue click Start Network Receive. This will present a file open dialogue window named Locate Central Server Config XML. Select the configuration file named CentralConfig.xml. Next another file open dialogue named Locate Calibration XML will be presented allowing the selection of the current calibration file.

After the calibration file has been selected connection to the capture nodes commences. Observe the command window, it should display an output similar to the following figure if all connections were successful:

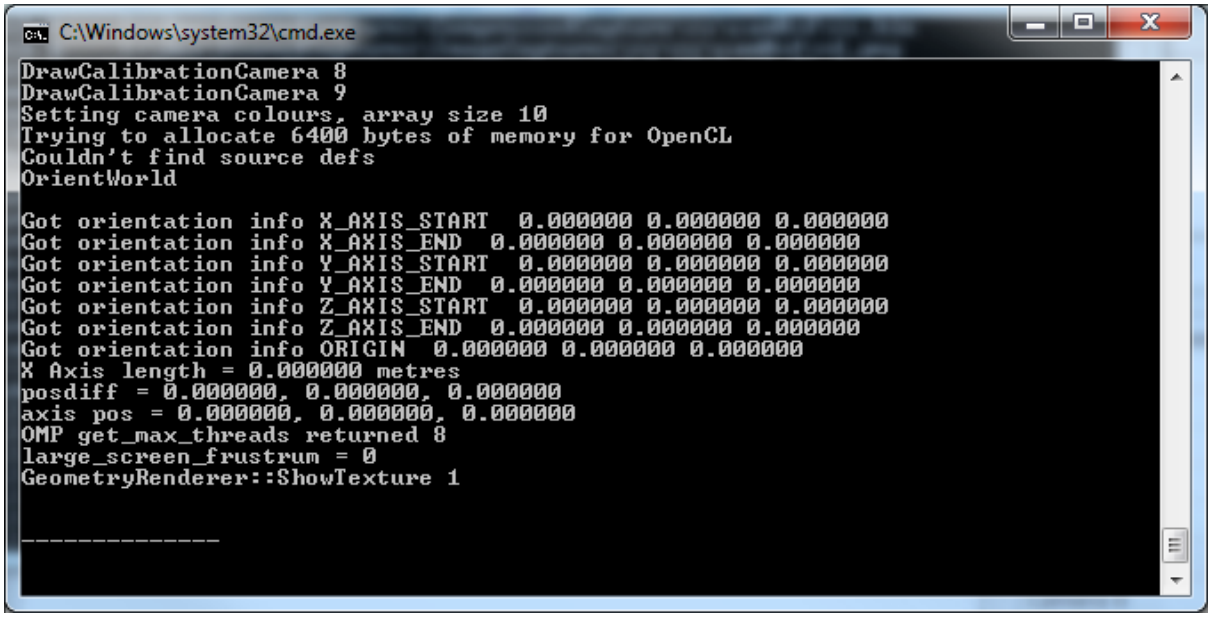


Figure 7 Successful Connection Output

Note: there is a deprecated option to stream from images that have previously been saved to disk on each of the capture nodes. This is done by editing the Central Server Config XML as follows:

```
<Capture device="PYLON" xmode="CONTINUOUS" xmode="SINGLE" showOnCentralNode="true"
showOnCaptureNode="true" isPush="false" />
```

Change the word PYLON to DISK.

```
<TestImages start="1" end="250" path="C:\Dev\TestImages\AlCapture300914b" />
```

Change the path to the images and the start and end frame number as required.

This also requires that the appropriate calibration file for the cameras (used when the image set was created) is loaded.

Loading image set from disk

To start the software in this mode execute the batch file named: '***run_with_geom_stream – no save.bat***' then once the 3D reconstruction software has loaded click the **'Start Geom Stream'** button in the *Camera Control* dialogue window that takes focus.

Important: do not forget to click the ***'Start Geom Stream'*** button.

3D Rendering

The rendering may be performed using the 3D Render Client software (Figure 8 3D Render Client).

The client can function in one of two streaming modes: live or parse from disk

Live

In this mode the client receives 3D mesh and video data from the 3D reconstruction node via a TCP/IP connection.

Parsing from disk

To parse a file from disk simply run software with the following command line arguments:

geom_stream_client.exe -o J:\ProtobufMsgs\ProtobufFileName.pb

Alternatively, for a more permanent and easier method of loading a particular file: edit and resave one of the batch processor files with a new meaningful name.

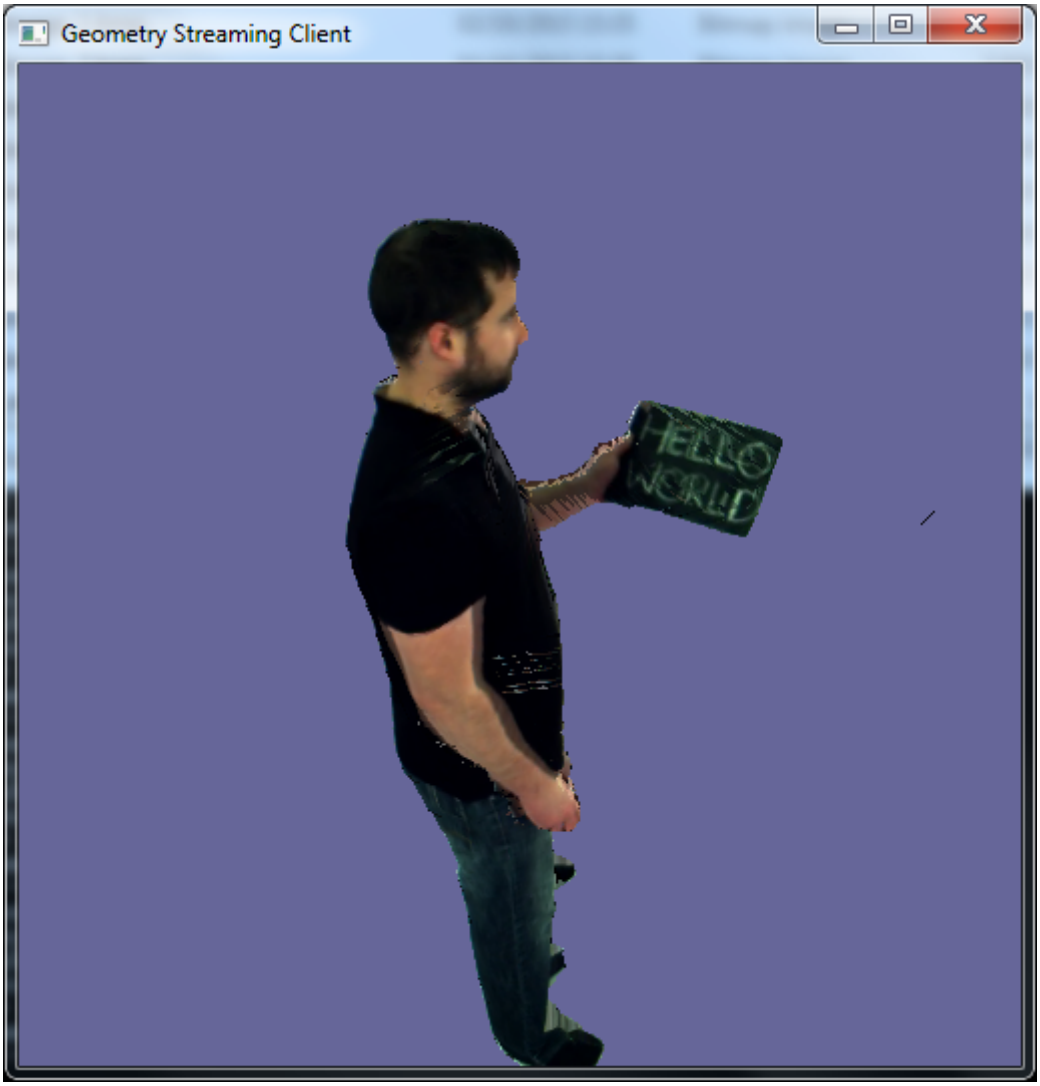


Figure 8 3D Render Client

Appendix B

A Mixed Reality Telepresence System for Collaborative Space Operation

Allen J. Fairchild, Simon P. Campion, Arturo S. García, Robin Wolff, Terrence Fernando and David J. Roberts

***Abstract*—This paper presents a Mixed Reality system that results from the integration of a telepresence system and an application to improve collaborative space exploration. The system combines free viewpoint video with immersive projection technology to support non-verbal communication, including eye gaze, inter-personal distance and facial expression. Importantly, these can be interpreted together as people move around the simulation, maintaining natural social distance. The application is a simulation of Mars, within which the collaborators must come to agreement over, for example, where the Rover should land and go.**

The first contribution is the creation of a Mixed Reality system supporting contextualization of non-verbal communication. Two technological contributions are prototyping a technique to subtract a person from a background that may contain physical objects and/or moving images, and a light weight texturing method for multi-view rendering which provides balance in terms of visual and temporal quality. A practical contribution is the demonstration of pragmatic approaches to sharing space between display systems of distinct levels of immersion. A research tool contribution is a system that allows comparison of conventional authored and video based reconstructed avatars, within an environment that encourages exploration and social interaction. Aspects of system quality, including the communication of facial expression and end-to-end latency are reported.

***Index Terms*—Computer supported collaborative work, mixed reality, telepresence, 3D video based reconstruction, background-foreground segmentation, space science.**

I. Introduction

THIS paper presents the integration of a telepresence system [1] and a Mars simulator [2], in support of a European Union funded CROSS DRIVE project [3]. CROSS DRIVE seeks to improve collaboration between countries across space mission control, science and engineering. The aim of the work is to support most Non-Verbal Communication (NVC) while contextualizing it both within a scientific simulation (of Mars), and a team of people "beamed" into it from different locations.

The motivation behind CROSS DRIVE is to reduce divergence in both planning and science that can creep in between the occasional expensive group visits to another country's simulation facilities. This would be simple if only technology was already available to support across a distance, the quality of dialogue achievable when a team is physically immersed together within a simulation.

Unfortunately, contextualizing a wide range of non-verbal communication within a simulation in which collaborators can move around, is difficult [1]. Approaches tend to favor either the range of non-verbal communication supported (video conference), the level to which its spatial contextualization can be communicated, or freedom of movement within the shared space (collaborative virtual environments). In simple terms, it is surprisingly difficult to communicate both what someone looks like and what or who she is looking at, without constraining movement, e.g. with seats. The problem is that both non-verbal communication and environment based problem solving are inherently spatial.

To understand the relevance of this problem to the CROSS DRIVE project, consider the following scenario. A scientist might point to where she thinks the Mars Rover should be sent. An engineer frowns and points first to the suspension of the Rover and then the terrain it would have to cross. However, seeing that only the mission controller is looking at her, she moves into the scientist's line of sight and throws up her hands. In video conferencing, what people are looking and pointing at would likely be lost and someone cannot walk into the line of sight of a remote user to capture attention. With immersive collaborative virtual environments using conventional motion driven authored avatars, facial expression and often identity would be lost.

The key challenge is to support a wide range of non-verbal communication contextualized within a real simulation and application. This paper describes a set of sub-challenges that were addressed. These include segmentation against backgrounds that may include moving images on a display, extending immersion of a wall display to allow another's space to be entered, real-time texturing of face without overly distorting its appearance, enabling scalability of a streamed 3D video avatar and sharing of spaces without occluding eyes.

Paper submitted: 09/10/2015. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 607177 CrossDrive. The UK's Engineering and Physical Research Council (EPSRC) through DTA and CASE PhD studentships, and through grant EP/E010032/1; and VISIONAIR through TNA-131 also supported this work.

A. J. Fairchild, S. P. Campion, A. S. García, T. Fernando, and D. J. Roberts are with the University of Salford, Salford, M5 4WT, U.K. (e-mail: a.j.fairchild@edu.salford.ac.uk; s.p.campion@salford.ac.uk; a.s.garciajimenez@salford.ac.uk; t.fernando@salford.ac.uk; d.j.roberts@salford.ac.uk).

R. Wolff is with the German Aerospace Center (DLR), Lilienthalplatz 7, D-38108, Braunschweig, Germany. (e-mail: robin.wolff@dlr.de).

II. Background

Mixed Reality (MR) merges information from the real and virtual worlds, using mediums and displays. Depending on the amount of virtual and real information, a particular application can fall in different points of the MR continuum described by Milgram and Kishino in [4]. Within this continuum, it is acceptable to place a Head Mounted Display that cannot be seen through in Virtual Reality (VR) and one that can in Augmented Reality. It is also understood that a natural environment overlaid by graphics is Augmented Reality and a virtual avatar abstracted from video of a user is Virtuality. It is less straightforward to place either the Mars simulator or its combination with our telepresence system within these discrete containers. Thus we describe it as a mixed reality system.

A. *Reproducing NVC in Telepresence*

Telepresence is the feeling of being in a different place derived from a technology. Technologies range from web cameras to embodied humanoid robots. The term is often used to describe systems that attempt to reproduce face-to-face meetings across a distance. Here we focus on such systems that range from video conferencing, through immersive virtual environments, to the combination of video based reconstruction and immersive displays. The focus is the affordances that each give to communicating NVC.

NVC has been described as the transmission of information and influence through an individual's physical and behavioral cues [7]. This transmission is usually via a range of cues that often only retain correct meaning when interpreted together and within context. There are different technologies that can be used in order to capture aspects of NVC for computer-mediated conversation. Video is sufficient for capturing most NVC within the filmed spatial/temporal context. However, when what is being responded to is out of view or delayed, the meaning of the response may be lost [1]. At the same time, it is difficult to capture mutual eye gaze, as the camera and display cannot share the same physical position in the space and the user can only look at one of them at a time [13]. On the other hand, VR, in its purer form, immerses people in 3D computer graphics, tracking some of their movements and hence capturing some of the NVC of the user. The use of immersive displays and life size motion tracked avatars make it possible to retain spatial context so the user can be seen by another remote participant sharing the environment [9].

Avatars of varying detail are used to represent users in Immersive Collaborative Virtual Environments (ICVE). The standard approach uses live motion tracking data from the user to mirror her movement through a remote avatar. This varies from simple head and hand tracking to more complex approaches, such as [8], incorporating eye gaze. This method of user representation has proven successful when completing collaborative tasks [9]. Studies have also illustrated how a number of NVC can be successfully portrayed using virtual characters [10]. However, capture and display of facial expressions in real time alongside full body tracking is a much bigger challenge. Affordable commercial software such as Faceshift [11] does allow for real time marker-less capture of facial expressions but relies on a depth camera being so close to the face to capture detail which would be problematic when also capturing the body. Marker based solutions could be used to animate a facially rigged character [12], however, this is not plausible for frequent use because it requires too much time for setup and could also make the wearer feel uncomfortable.

A contemporary technology approach to telepresence is the use of 3D reconstructed video for communication [5][6]. In such systems, avatars are created in real-time from several video streams. The road that led us to such an approach was building and comparing the use of gaze enabled ICVE and video conferencing [A]. Each offered different affordances to NVC which meant that each only told part of the story [1]. In order not to confuse the story, we needed to faithfully communicate at least eye gaze, interpersonal distance and facial expression [1]. The advantage of combining video based reconstruction and immersive displays is that it attempts faithful transmission of all of these and more [1]. If people are going to act naturally, it helps if they are encumbered by excessive markers and restrictions on movements within extent of social space [1]. Another advantage of the above approach is that the only thing that needs tracking is viewpoint [1] and this can be done across extent of social distance by, for example, placing markers on glasses or a hat. However, a challenge is supporting a sufficient balance of visual, spatial and temporal quality [1].

B. *Live generation of 3D avatars*

Live generation of 3D avatars can be achieved actively or passively. Active methods include time-of-flight devices that project light towards and analyze the time it takes to reach points on an object [14], and structured light devices that analyze disparity in a projected pattern to form a 3D representation [15][16]. The Kinect is an example of a structure light device that has been used in much recent telepresence research. A single Kinect can achieve a partial 3D reconstruction of the subject in the plane it is pointing towards. However, full 3D reconstruction [1] is required to allow people to use movement in space as part of communication, so that a person does not look like an empty shell when viewed from the side. This requires the stitching together of depth maps from multiple Kinects positioned around the subject. Herein lays a problem, because the projected patterns from the individual Kinects interfere with one another, and this causes deterioration in the quality of the depths maps, typically resulting in less faithful shapes with holes in them. Interference between multiple Kinects can be reduced [17] and there are numerous examples of Kinects being used for 3D capture [18][19][20] but, to the authors knowledge, only two produced a 3D avatar that was generated without the surrounding environment [21] [22]. However, there is a bigger problem. The resolution of structured light patterns on a face is far less than that of pixels capturing a face from the same distance with an RGB camera [1]. This results in poor resolution of shape of face if cameras are far enough away to allow natural interpersonal distance [1].

With passive methods, also known as Image Based Reconstruction (IBR) [23], the 3D model is derived from a set of conventional camera images taken from different angles and positions, capturing light in the visible spectrum. These methods then use this information to generate form (geometry) and appearance (texture). Video Based 3D Reconstruction (VBR) extends this across time to also capture movement from multiple video streams. There are several VBR approaches suitable for reconstructing the 3D form of an entire human that fulfill the requirements of our system. One example is multi-view stereo [24], which is capable of producing high quality, spatially accurate and visually faithful models. Unfortunately, it currently falls short of the temporal requirements of a real time telepresence system. Techniques based on the shape-from-silhouette (SfS) principle [25], which form an approximation to the 3D shape known as the visual hull [26], have demonstrated that they can fulfill this requirement whilst retaining a faithful reconstruction [1]. For that reason, methods to extract silhouette information required for SfS become paramount.

Both active and passive methods have strengths and weaknesses when applied to 3D telepresence avatar generation. Multiple Kinect based approaches are currently of a lower resolution compared to that which can be achieved with SfS using conventional cameras with resolutions typically in excess of 1000x1000 compared to 320x240 pixels depth map resolution. They offer a less faithful reproduction because the holes produced due to pattern interference need to be filled and what fills them may not be a true representation of the real world. Moreover, there is a drop in quality of depth maps over distance [27] and this reduces the potential capture volume, which is not desirable for user movement or interacting with objects. SfS currently offers higher textural resolution and does not suffer from holes thus enabling clearer representation of eye gaze and facial expressions both of which are vital for portraying accurate NVC. Depth based approaches, however, can be deployed within an immersive environment where as, with the exception of [28], SfS requires a sterile background that would prevent the system to fully immerse the user.

1) Texturing

After capturing and reconstructing the user in 3D, texturing is needed to provide a life like representation. A composition of the segmented images of the different cameras is then used to generate the final texture. Our previous approaches to multi-view rendering used these images with no blending [13] and this resulted in a clear eye and face representation, but with undesired visible lines in the border of the different images. The use of image blending [29] improved the quality as there were no visible lines, but this process blurred the eyes, limiting the set of NVC that the reconstructed avatar was able to convey. Floating Textures [30] is a method that provides high quality results, however, it is complicated, and we found no evidence published of the quality of the eye reproduction. This suggests that a simple yet sufficiently effective method can be used.

C. Contextualizing a wide range of NVC by combing immersive displays and live reconstruction

Non-verbal communication is inherently spatial. Retaining this spatial context is highly challenging with today's displays and mediums. In particular much thought has to go into the way in which medium and display are combined. Both the medium and the display impact on the way in which space can be shared. This in turn impacts on the contextualization of NVC. Both [1] and [31] allow users to view each other but not physically walk into each other's space. The former shows the remote user within their space, whereas the latter shows them in a simulated space.

Immersive displays can vary from HMDs to large immersive projection-based displays. Unfortunately, not all of them are appropriate given the requirements of the system proposed. HMD's have been combined with video based reconstruction [ref] but this completely hides eye gaze. Immersive projection technology usually uses 3D glasses, which at best make eyes hard to see [31]. In the closest studies identified attempting to support collaborative meetings, two users, captured using a single Kinect [32] and two Kinects [33], were reconstructed in front of a collaborative whiteboard, allowing visual communication of NVC and written notes from a fixed perspective. Our study uses 10 cameras to capture the user, and thus it is possible to walk completely around them while they are contextualized within a much larger synthetic environment, in a similar way to [34].

1) Foreground segmentation from a background containing moving objects or images.

Sharing a completely simulated space requires that people but not their surrounds are transmitted. Different approaches to background-foreground segmentation include simple background subtraction [35], Chroma Keying [36] and more advanced background-foreground detection methods such as [37][38][39]. The choice of background-foreground segmentation method impacts on the faithfulness of the reconstruction.

Our initial approach to combine immersive projection technology with free viewpoint video was inspired by the BBC [40]. This used a retro-reflective material to allow the user but not surrounding cameras to see a projected image. However, this proved to have a number of drawbacks. Firstly, the material does not allow projection from the rear. The projection quality is lower due to material properties. Lastly, the retro-reflective qualities of the material require many projectors to support viewing across a typical display volume. We then developed a solution that segmented a background of unified color **[B]**. However, this limits the user to looking into rather than sharing other's space, as in [1] and [31]. Another consequence of the need for a unified background color is that the solution is not readily deployable to most simulation facilities. It is desirable to be able to subtract backgrounds comprised of both static objects and moving images. The work toward a solution is described later in this paper.

III. Collaborative Mixed Reality System

The collaborative MR system is realized via combination of enhancements to an existing telepresence system [1] followed

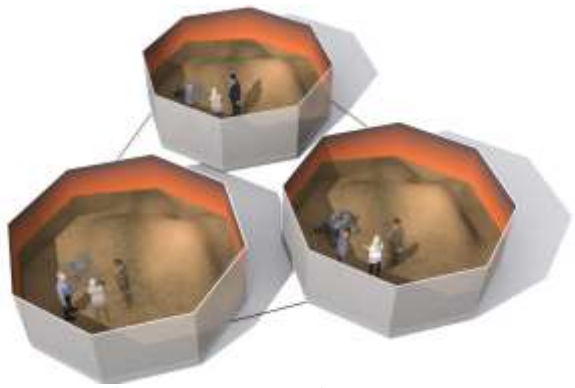


Fig. 1. Illustration of how the ideal system for three users would look like: every user would have one immersive display and all of them would share the same virtual space where the others are faithfully reconstructed in 3D.

by its integration with the Mars simulator. First, this section presents an ideal immersive projection telepresence system, and then the telepresence and Mars simulator systems developed to date, and finally their integration.

A. *Ideal immersive projection system*

The following system would require that the most advanced display and capture equipment was available at all sites. With hardware extended to support real time segmentation of users against live simulated backgrounds across the entire volume of each display. It would also feature stereo display enabling eye gaze clearly visible at three meters. Finally, a variety of the multidimensional datasets, with different spatial and temporal resolutions, would be available for Mars, including Rover simulation capabilities.

Using this system, each user would be able to move around the Mars simulation within the extent of social space with others, whether in the same or distributed locations. Enabling natural movement around, for example, a Mars Rover, bringing users' attention to attributes of it and the surrounding environment. Fig. 1 depicts this ideal situation.

B. *Current systems*

This section describes the current state of both the telepresence system and the Mars simulator paying special attention in the updates carried out to meet the requirements of the MR system.

In the context of CROSS DRIVE, there is only one fully immersive display available, the octave [41], the rest are Powerwalls and desktops computers. This results in only one user being able to walk 360 degrees around an object while still retaining eye contact.

Furthermore, the octave is the only one equipped for multi-camera video capture. However, segmentation of the projected simulation is currently not supported across the entire space. This restricts projection of simulation to a single wall but outside of the view of the cameras.

1) *Telepresence system*

An update on the 3D telepresence research system *withyou*

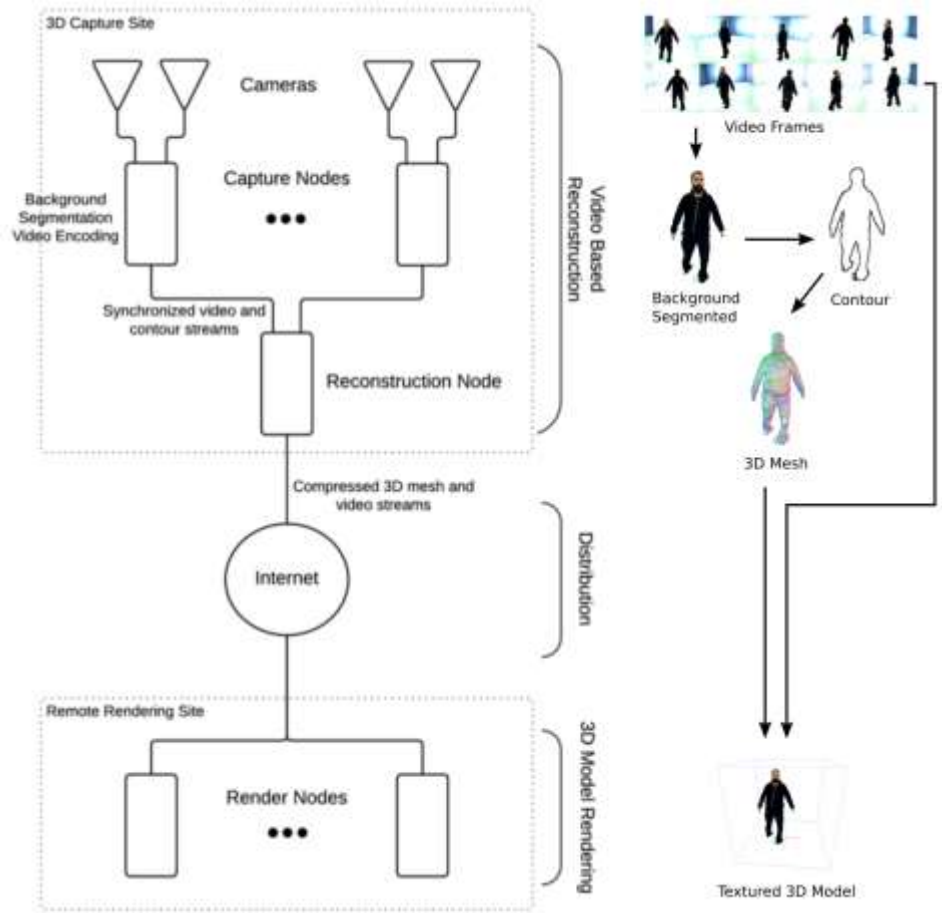


Fig. 2. Generic architecture of our telepresence system.

[1] is presented in this paper. First, the end-to-end system architecture is detailed followed by a description of extensions, with justifications, that have been made.

The complete end-to-end system architecture is comprised of multiple network connected components with each contributing to the processing pipeline that is originally described in depth in [1] and summarized in Fig. 2. This figure depicts the high level visualization of the telepresence system architecture from the cameras that acquire the subject(s) (top) through to rendering on the end nodes, which can be in different and possibly geographically dispersed locations (bottom). The following subsections outline the whole process.

a) Video Based Reconstruction

The outcome of this first stage of the process is the 3D mesh that represents the user together with the video streams that will be used to texture it. This includes the subject acquisition, the background-foreground segmentation and the 3D model generation.

(1) Image Acquisition

The process begins with the acquisition of images of the subject(s) via an array of cameras surrounding them. Cameras are either mounted on tripods or above the displays depending on the display configuration.

(2) Background-foreground Segmentation

A particular challenge of this work is that a user may be stood against a background containing static or moving images. Currently, there are two implementations of segmentation, the first is fully implemented and the second partially.

- Segmentation in the visible light spectrum: In previous publications, the system utilized a GPU Mester based background-foreground segmentation method [42]. However, the new requirement of a more faithful 3D reconstruction posed by the CROSS DRIVE project was not met. Also from a practical perspective, the previous method required domain specific knowledge to configure each time a camera was repositioned so experimentation with different setups [13] was a painstaking process. To improve the faithfulness of the avatar and alleviate the configuration constraint, the system was enhanced with a GPU implementation of Gaussian Mixture-based segmentation [39]. The shadow detection [43] has also assisted especially round the feet of the user. Moreover, in the sterile environment of the octave, it requires no domain specific knowledge to configure, thus enabling researchers to change camera positions without reconfiguring.
- Segmentation in the infrared (IR) light spectrum: Creating a 3D reconstruction of a user while she is immersed in an environment with a moving background is challenging. Regarding the segmentation in the octave, one of the problems with VBR in a fully immersive display system is that the user is surrounded by the display and thus the segmentation method employed needs to extract their silhouette from a moving background. As has already been commented in Section 2, this posed a limitation in the octave resulting in one screen being used for the display, reducing the expressiveness of the 3D avatar (the user was not able to point or look at things out of the screen). In an attempt to solve this limitation, a solution utilizing the IR light spectrum to perform segmentation is being considered. The subject stands in the middle of an immersive display and is illuminated with IR light from strategically positioned surrounding lamps. The user and surrounding background are acquired by cameras that are only receptive to light in the same frequency as the light emitted by the lamps. The cameras only capture the objects illuminated in IR light and nothing projected on the screens, thus, the moving background no longer interferes with the segmentation. The IR camera is physically positioned in close proximity with a visible light camera pair and its pose in relation to it determined using the checkerboard calibration technique.

The results of these two approaches are presented in Section IV where their impact on the quality of the 3D reconstructed avatar and the combination of 3D reconstruction and immersive displays is shown.

(3) 3D Model Generation

Upon receiving and decoding the video streams and contour data from the capture nodes, the reconstruction node generates a 3D model avatar via a parallelized "Exact Polyhedral Visual Hulls" (EPVH [26]) implementation [44]. To generate a 3D model the system requires knowledge of the cameras image planes in relation to real word 3D coordinates [45].

b) Distribution

In the *withyou* system, both model generation and rendering were executed on the 3D Reconstruction node and this limited the practical usability of the system. To overcome this restriction, a new method of distributed rendering was proposed and implemented. The rendering process was detached from the 3D reconstruction component and placed in its own self-contained client. To allow for multiple remote rendering sites, the new renderer is network enabled. After the 3D model generation, the reconstruction node prepares the 3D mesh and video data for broadcasting to connected remote rendering sites, packaging all relevant data into a network message. A message contains vertex positions, triangle indices, a video frame per camera, as well as frame number and timestamp. In order to reduce the amount of data sent across the network, the 3D mesh is compressed using the LZMA algorithm [46] after serialization, and this results in between 67% and 75% reduction in size. The h.264 encoded video is taken directly from the input of the capture nodes to avoid decompression and recompression by the reconstruction component. Synchronization of the video and mesh data is handled by placing the data together in the same network message.

c) 3D Model Rendering

A new texturing method is implemented with the aim of removing visible lines at polygon joins without confusing the image through blending. Another goal was to test if this could be achieved with an approach simpler than [30].

The render node decompresses the incoming geometry

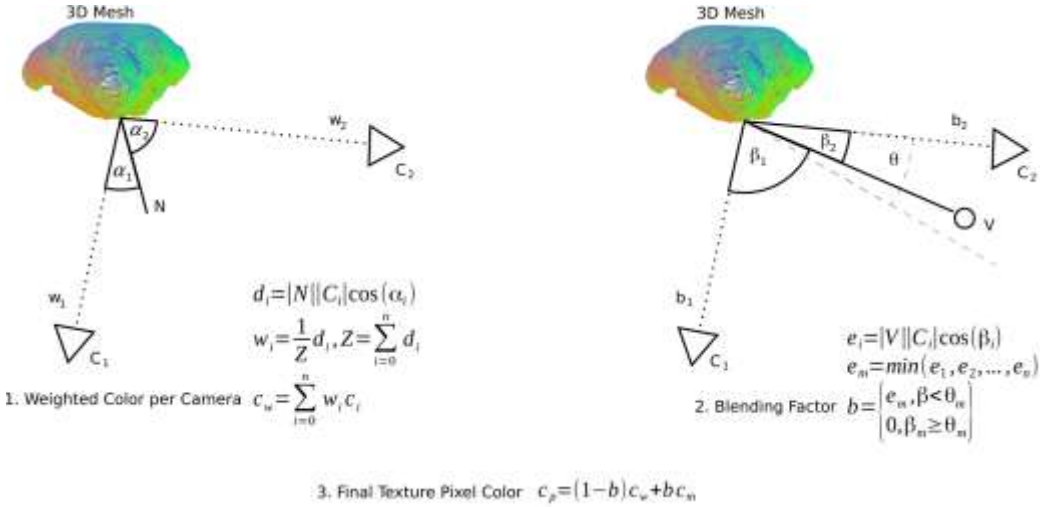


Fig. 3. Computing the final pixel color of the texture by combining weight based blending with viewpoint dependent blending of projected camera images. 1.) Weighted blending based on the angle between surface normal N and camera vector C_i, where w_i is the weight of a camera, c_i is the projected pixel color of a camera, and n is the number of cameras. 2.) Viewpoint-dependent blending based on the angle between the viewpoint vector V and vector of the closest camera, where e_m is the smallest angle and θ is a threshold. 3.) The final color is the combination of the weighted color and the color of the closest camera to the viewpoint blended-in, where c_p is the resulting pixel color, c_w is the computed weighted color of a camera, c_m is the projected pixel color of the camera closest to the viewpoint.

mesh data and computes vertex normals via the weighted average of the angle between connected triangle edges. It then pushes the vertex positions, normals and triangle indices into OpenGL buffers on the GPU. The compressed video frames are decoded and pushed directly onto texture buffers on the GPU.

Texturing is realized in a pixel shader program in which each texture is projected onto the mesh from the corresponding camera perspective. The algorithm computes the color of a pixel based on a weighted blending of projected pixels from the camera images. The blending weights w are determined by computing the dot product between the fragment normal N and the direction from the fragment to a camera C, so that $w = 1$ with $\alpha = 0°$ and $w = 0$ with $\alpha \geq 90°$. The weights are then normalized so that their sum equals one and applied when adding the projected pixel colors of the respective camera images, see Fig. 3, 1.

The weighted blending method provides that surfaces facing closer toward a specific camera receive a higher contribution to the final pixel color from this camera's image than from others. The result is a smooth blending of the projected textures. While this method is simple and does not require on a specific camera arrangement, it can cause distortions in areas without a dominant camera and where cameras have similar weights. Furthermore, it does not take occluded areas into account.

In order to further improve visual quality, the texture mapping algorithm has been extended with a viewpoint-based blending method, where a camera image that was captured from a direction close to the current viewing direction of the user has higher influence than the color determined via the surface normals as described above. The algorithm starts with finding the closest camera by comparing the angles between the camera directions (vector from surface to camera) and the direction to the current viewpoint (vector from surface to viewer), see Fig. 3, 2. If the smallest angle is below a threshold, then the image of this camera is blended over the previously computed texture, see Fig. 3, 3. The blending factor is inversely proportional to the angle between the closest camera and viewer direction and ranges from zero to one.

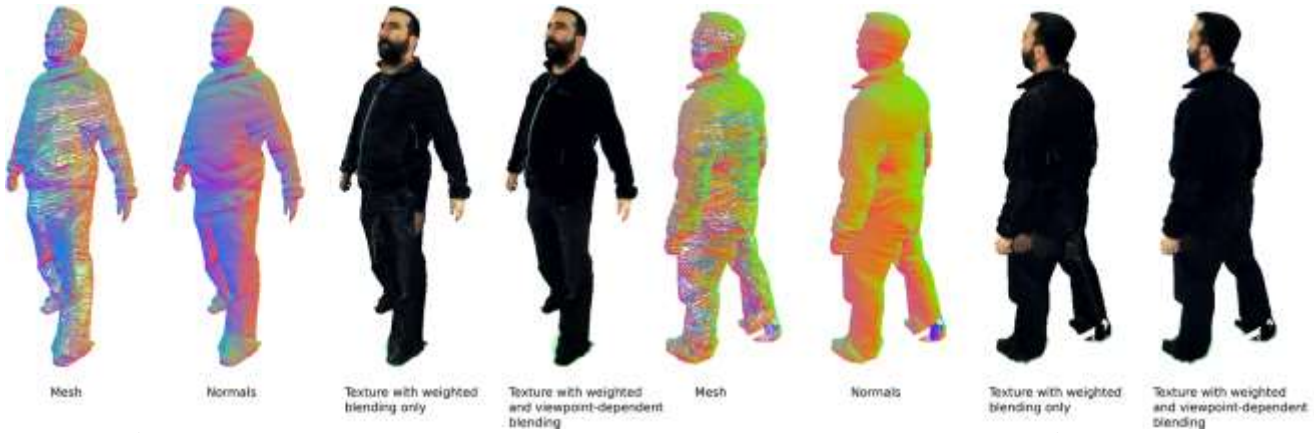


Fig. 4. Stages of 3D model rendering: incoming mesh; normal generation; texture generation via weighted blending of projected camera images; and blending of the image of the camera that is close to the user's viewpoint.

Smaller angles produce a higher blend-in factor and an angle of zero results in fully displaying the pixel of the closest camera.

A suitable choice for the threshold is influenced by the arrangement of capture cameras and the preference of blending behavior. A narrow threshold causes the texture to fade-in only when the viewer is very close to a camera view, whereas a large threshold causes the texture to fade-in from a larger distance. In our setup, a threshold of 12 degrees was chosen.

The result is shown in Fig. 4. The combination of both texture mapping methods provides the best compromise of computation effort and visual quality for our system. Although, the viewpoint-based blending technique is only effective when the viewer looks at the reconstructed mesh from near a camera view, it significantly improves the visual quality at these occasions. For example, the collar is correctly colored in the front view and the ear is rendered with a shadow, see Fig. 4. As our texture mapping technique does not test for visibility of surfaces to cameras, the viewpoint-based blending method has a further advantage, as it hides wrongly applied pixels to occluded areas. For example, with the simple weighted blending method based on surface normals, the pixels of the hand captured by the camera to the left of the subject are mapped onto an area of the reconstructed mesh (near the hip), see Fig. 4. By applying the viewpoint-based texture mapping method, the image of the front camera is blended over the weights-based texture and the projected hand on the hip disappears.

2) Mars simulator

At this stage of the CROSS DRIVE project, only geology datasets have been integrated into the Mars simulator, making it possible to study the surface of Mars with different digital terrain model (DTM) resolutions and even using subsurface data obtained from subsurface sounding radar.

Therefore, the simulator is based on a VR cartography system designed for interactive exploration and analysis of a planet's surface within immersive virtual environments [2]. The system is capable of visualizing very large DTM datasets at interactive frame rates while assuring that the best available resolution is always shown. The renderer creates DTMs from geo-referenced raster data and provides interactive tools commonly found in Geo-Information Systems (GIS).

C. Integration

The Mars simulator provides interactive geology tools and supports collaboration between remote people within the immersive virtual environment using traditional authored avatars. In this paper, it is used to explore the surface of Mars. Each participating site runs an instance of the simulator with the Mars data stored locally. User interactions are then synchronized by a collaboration manager created within the CROSS DRIVE project. In its default configuration, the instances use a traditional CGI character as an avatar to represent remote users in the virtual environment. The Mars simulator has then been extended for supporting 3D video avatars. The extensions include a communication module for receiving the 3D mesh and video stream, as well as a rendering module for visualizing the 3D reconstructed avatar within the Mars terrain renderer.

When a participant joins a collaborative 3D Mars exploration session from inside a 3D capture space (octave), then this user will be represented by a 3D reconstructed avatar instead of the traditional CGI avatar. The system communicates the server address of the reconstruction node to the other participants, which open a connection to the reconstruction node directly for receiving the 3D mesh and video stream. The server starts streaming as soon as a remote client is connected.

The communication module runs decoupled of the rendering in a separate process, whereas the rendering module is triggered each rendering frame within the Mars simulator. If new data is available on receiving sites, the data is copied from the network message buffer; the mesh is uncompressed and the video decoded; and the rendering process, as described in Section B.1)c), is initiated.

The position of the local user watching the 3D reconstructed avatar, needed for the viewpoint-based texture mapping, is provided by the head tracking system of the immersive virtual environment.

Additionally, a transformation matrix has been added that, firstly, aligns the origin of the capture space, and thus the origin of the 3D reconstructed avatar, with the origin of the virtual world in the 3D Mars simulator, and secondly, scales the units in the capture space to match the units in the renderer space. This way, when moving around and pointing at references within the 3D Mars simulation in the capture space, the reconstructed avatar appears in life-size and at the corresponding position and orientation within the simulation in the remote virtual environments. With our current camera configuration we can capture and reconstruct people so that their gaze and facial expression are clear while they occupy any position within 1.5m radii from the center of the octave.

IV. RESULTS

This section overviews the qualities of the system. It provides evidence toward validating the approach, although neither a perceptual or behavioral study is provided. However, we hope it provides sufficient evidence that such in-depth studies would now be achievable. We argue that our balanced approach to supporting interpersonal movement, gaze, facial expression and the integration of an application to encourage their use, opens the door to such experiments.

A. Visual quality and communication of NVC

Firstly, previous and new methods for segmenting in the visible light spectrum are compared.

Fig. 5 shows the finer granularity achieved with the new approach used. Notice less jaggedness and closer match to the actual form of the face (indentation at bridge of nose, mouth, hairline and chin) and better representation of the digits of the hand without webbing effect. This finer granularity results in the generation of more faithful avatars.

With this improvement, we have been able to demonstrate

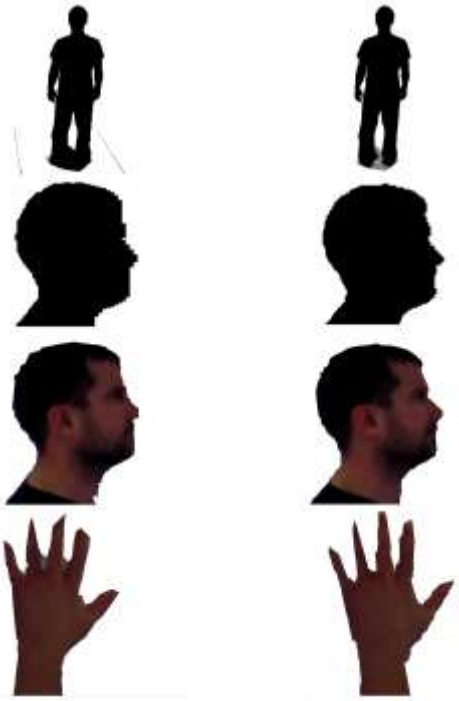


Fig. 6. Segmentation comparison of previous MESTER (left) and current Gaussian Mixture-based (right) implementations. Notice incorrectly classified regions and jagged edges that were present using the previous method (left) compared to the finer granularity, including shadow detection highlighted in grey, of the current method (right).

the systems capability to show gross NVC such as waving (Fig. 8), pointing and interpersonal distance (Fig. 6). In addition, the system is capable of capturing and displaying subtler NVC such as eye gaze and facial expressions. Fig. 7 shows the quality and clarity of facial expressions achievable with a good camera calibration. It illustrates the seven universal emotions described in [47]. Highlighting that the reconstruction quality is high enough to achieve this and in addition it shows quality of eye gaze captured. It should be noted that camera rig height can have an impact on reconstruction quality inducing a droop effect that can make a user appear sad, aged or unwell, which becomes worse as the user approaches the outer limits of the capture space [1]. The camera setup for Fig. 7 appears to be set just right for the user being captured yet the cameras are all above the screens.

Apart from the quality of the reconstructed 3D model, the fact that users may have different hardware available to them can raise new issues. Fig. 9 shows three images of a 3D reconstructed avatar of a user. This highlights the issues of different display technologies when used with 3D reconstruction. If shutter glasses are used to enable stereo, then eye gaze is not captured and the facial muscles around the eyes are partially occluded. If a full HMD is used, then very little of the face is visible. The result is a complete lack of facial expressions. The system described in this paper is capable of utilizing both as we recognize that not every user will have a 3D capture system or immersive display.

B. Segmentation from static or moving background

A specific goal was to segment against background including moving images. This section shows the preliminary results of a new approach in IR light spectrum to allow segmentation across immersive displays and static backgrounds. For this proof of concept, we used Kinect as an IR camera, but it will be replaced for higher resolution cameras in further tests.

We have confirmed that both projection systems in the octave and one of the Powerwall are not emitting IR light at a frequency that interferes with the Kinect's IR camera Thus, it is possible to segment the user against a moving background with it. Although not tested, we have no reason to assume that the result in the other Powerwall would be different. Fig. 10 shows the preliminary results of experimenting with different IR emitters and lamp positions and the effect on segmentation results due to differences in scene illumination. It is clear that the result is best when two IR lamps are active (as shown in the bottom-left part of Fig. 10). Further experimentation is now required to determine if the addition of more IR lamps and perhaps cameras with greater resolution than the Kinect could improve the result.

C. Temporal quality

A quantitative temporal evaluation of the existing *withyou* platform is presented in [1]. As a summary, the time taken to acquire a sequence of frames from 10 cameras then segment,

Fig. 5. Ability of the system to convey pointing gestures (left) and interpersonal distance (right).

Fig. 7. Universal facial expressions of emotion and eye gaze. This figure shows a 3D reconstruction of an author attempting to display the seven universal emotions and quality of eye gaze in three directions.

TABLE I
DISTRIBUTION AND RENDERING SUBSYSTEM TIMINGS

Process	Time (ms)
Mesh compression	118.5
Mesh decompression	23.4
Video decompression (10 frames)	16.1
Upload textures to GPU	105.2
Upload mesh to GPU	27.3
Render	0.05

encode video and reconstruct the 3D mesh is 79.32ms. This section presents some results of the new distribution and rendering subsystem (Table I) followed by the end-to-end latencies observed during the linkups.

The first stage is mesh compression, which is currently achieved using the LZMA algorithm. This is followed by serializing the compressed mesh, timestamp and encoded video frames. Table I presents times for processes that are required to distribute and render the participants at the remote sites.

The packet format and sizes are shown in Table II. The packets are then distributed to the render clients via network connection. Upon receiving the packets, the video frames are decoded and the mesh decompressed.

Whilst conducting the linkup, the authors carried out some latency tests for the system and repeated with Skype for comparison. The results are shown below in Table III.

The update to the segmentation procedure described in Section III.B resulted in an improvement in the accuracy and thus in the 3D form of the reconstructed avatar. However, it is currently hindered by our aging hardware. As a consequence, a reduction in the temporal quality of the system has been experienced. With newer hardware the temporal issue could be

TABLE II
TYPICAL PACKET COMPONENT SIZES

Item	Size (bytes)
Calibration data	740
Encoded video frames	9258
(10 cameras)	(average of 100 frames)
Compressed vertices	84745
(average number of vertices 9214)	(average of 100 frames)
Compressed triangles	47266
(average number of triangles 52607)	(average of 100 frames)

TABLE III
LATENCY OBSERVED DURING THE LINKUPS

	Local (Octave Salford)	THINKlab (Salford)	DLR (Germany)
3D Reconstruction	1.06s	1.12s	1.5s
Skype	-	0.103s	-

resolved thus enabling overall improvements. Table III demonstrates this by presenting the mean time taken to segment a sequence of 50 images on a number of different GPUs. The GeForce GT 730 is the card currently deployed in the reconstruction system.

Another aspect influencing the temporal qualities of the system is the streaming of the reconstructed avatar due to the high requirements on bandwidth. This is due to the fact that the 3D geometry and several HD video frames are streamed across the network. Although simple mesh compression and fast h.264 video are used, relatively low framerates are currently achieved. In order to avoid sending large amounts of data to several remote users, a proxy server was set up at the site in Germany. This reduced the traffic to a single stream from the UK to Germany and allowed us to distribute the 3D video stream to users inside the LAN and to the cluster nodes of our multi-pipe visualization system. However, this leads to increased delay. More advanced compression and data reduction methods are necessary to reach high framerates.

D. Initial linkup test

This section summarizes the initial connection tests of the system across Europe.

In this linkup, the surface of Mars is explored using elevation and imagery data from NASA' Mars Reconnaissance Orbiter (MOLA data, 500m/pixel) and ESA's

TABLE IV
COMPARISON OF SEGMENTATION TIMES WHEN PROCESSING TWO STREAMS SIMULTANEOUSLY USING VARIOUS GRAPHICS CARDS

GeForce GT 730	GeForce GTX 660	Quadro K5000	GeForce GTX 970
27.17ms	5.79ms	5.31ms	2.82ms

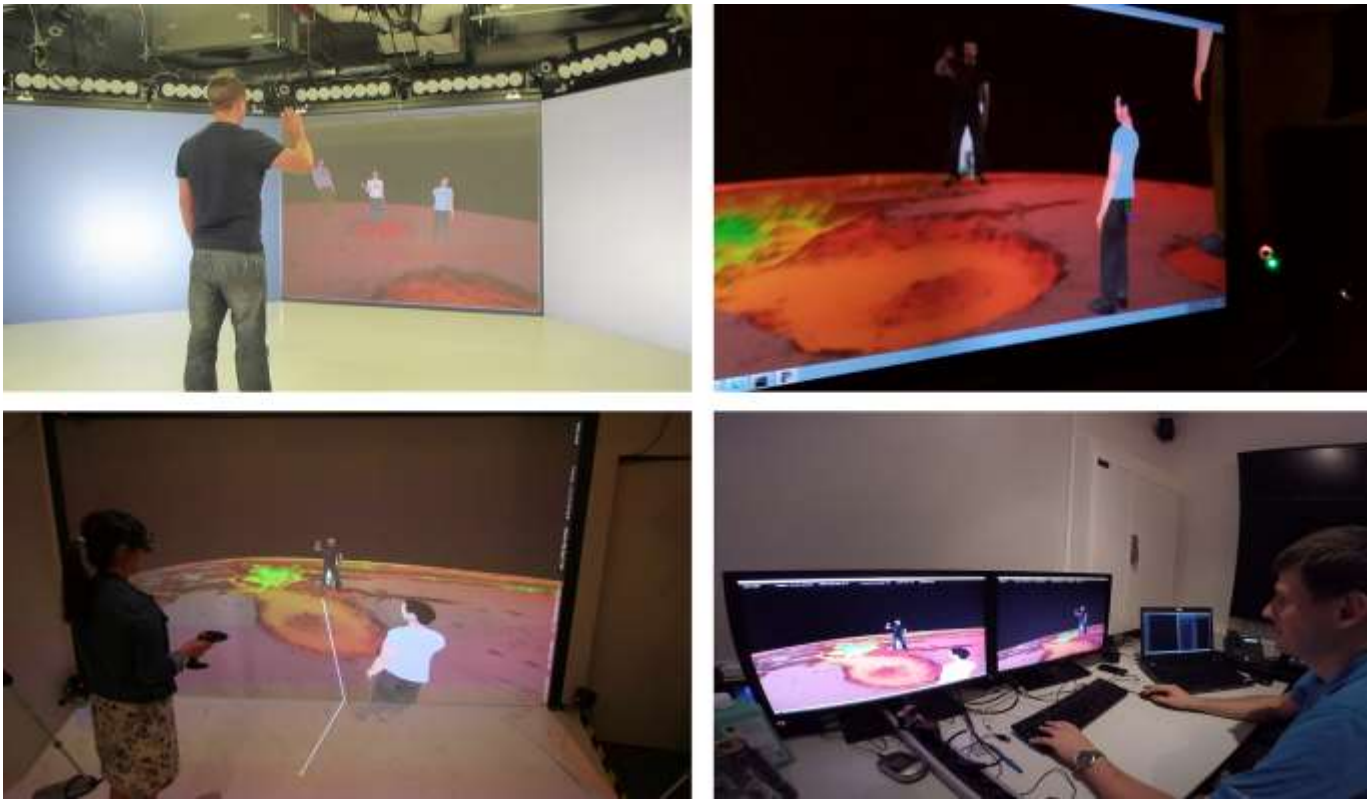


Fig. 9. Users greeting each other from Mars. This figure shows a user waving at three other users and his 3D reconstructed avatar from the different sites that took part in the experiment. The top left picture shows a user in the octave waving at three traditional avatars. The top right picture shows the 3D reconstructed avatar waving in the ThinkLab´s Powerwall (UK). The bottom left and right show this action as it was viewed by the Powerwall and the desktop system in DLR (Germany). The stereo view was removed from the two Powerwalls to take the pictures.

Mars Express (HRSC data, 12m/pixel) with a data volume of more than 600GB.

In contrast to the ideal immersive projection system depicted in Fig. 1, the current system deployed for this initial test does not feature immersive systems that surround the users. Fig. 11 illustrate this current configuration that shows three users, only one of them is 3D reconstructed (the person in the octave) while the others are represented by traditional motion-driven avatars.

Fig. 12 shows the components and interconnections of the three sites that were linked in the test, including the configuration of the capture system at the octave. Four different users took part, over three sites, two connecting from each country (see Fig. 8). Since only one site is currently able to generate and stream 3D video avatars, the rest of them were represented by traditional authored avatars. The second site, also in Salford, was the ThinkLab, using a stereo Powerwall and an optical tracking system. The third site, DLR (Germany), had one user connected using a stereo Powerwall with floor extension and optical tracking system, and the other using a desktop system.

Fig. 8 shows four pictures showing users greeting each

Fig. 10. This figure demonstrates the problem of occlusion (for the viewer) when using various display devices. Left no stereo (full facial expressions), center stereo glasses (eye gaze and some facial features obscured) and right HMD (Most of the face obscured, identification of facial expressions not possible).

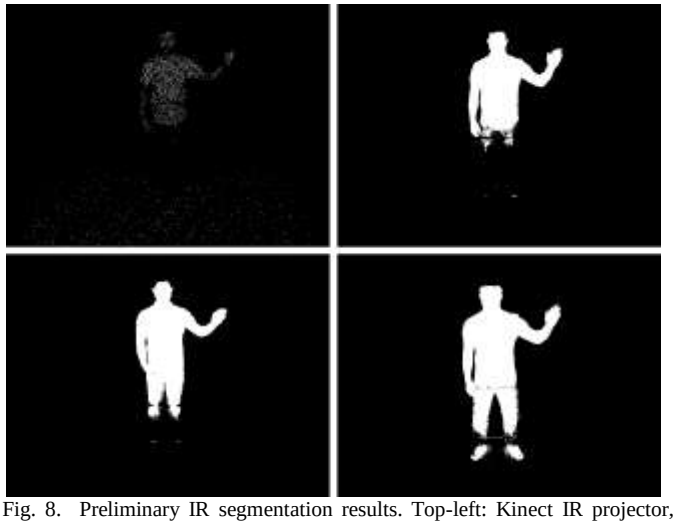


Fig. 8. Preliminary IR segmentation results. Top-left: Kinect IR projector, top-right: IR lamp mounted aligned with Kinect, bottom-left: IR lamp positioned on floor angled upwards and bottom-right: both IR lamps.

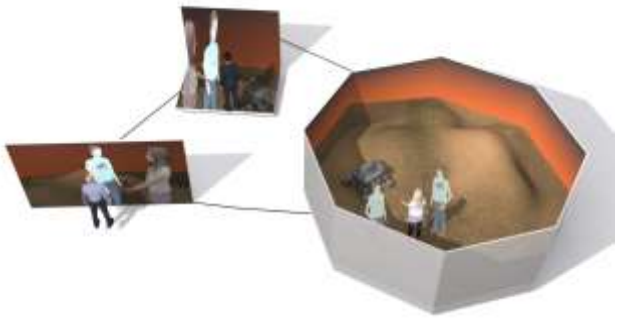


Fig. 12. Illustration of the current system: only one user has a fully immersive display and she is also the only one faithfully reconstructed in 3D. Note: The system tested utilized only one screen of the octave however the diagram reflects what is achievable with the new segmentation method.

other from Mars. The first picture shows a user waving at the rest from the octave, whereas the other three show this action performed by the 3D reconstructed avatar through the viewpoint of each of the other users. The participants could freely navigate and talk to each other. Fig. 13 depicts how the 3D reconstructed user is viewed from different viewpoints as one user moves around him. This free viewpoint navigation offers the user interactive functionality over and above that offered by traditional video conferencing, allowing users to

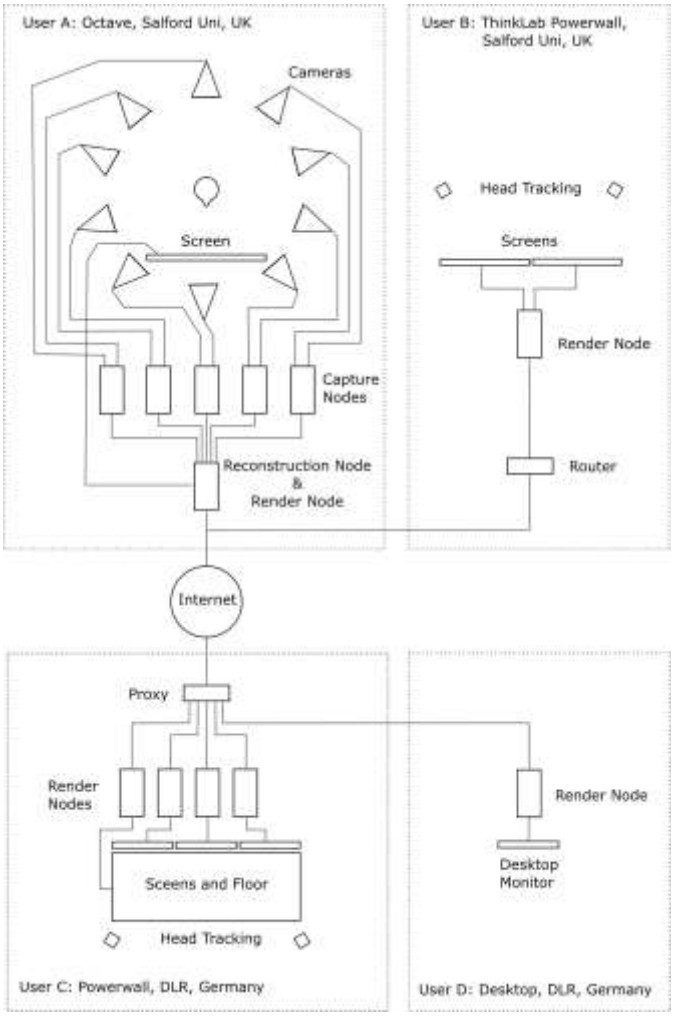


Fig. 13. Linkup architecture. This figure shows the current architecture of the initial test carried out.

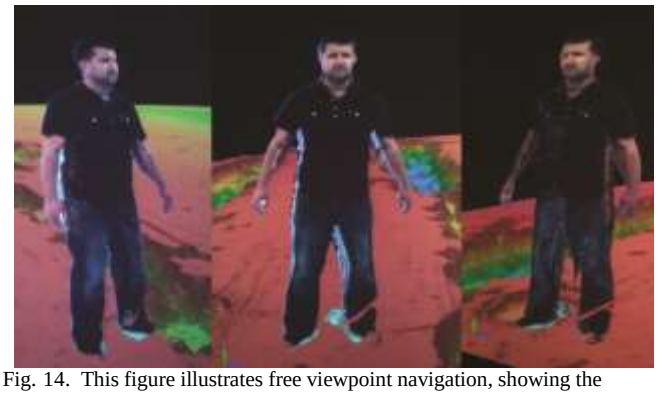


Fig. 14. This figure illustrates free viewpoint navigation, showing the capture subject as viewed from the left, centre and right of the display.

view NVC from any angle.

The 3D avatar is streamed from the octave, but instead of sending it directly to each user, a proxy was used in DLR to send it to the desktop client, reducing the bandwidth needed in the octave.

V. Discussion

This paper presented the integration of the *withyou* telepresence system and a Mars simulator that will allow scientists in remote locations to collaborate whenever necessary (saving travel time and money), while simultaneously exploring data sets. The former is designed to facilitate natural interplay between interpersonal distance, interactional eye gaze and a representative range of non-verbal signals (including facial expressions) associated with emotion, familiarity and trust. The latter provides a shared context and application where people in remote spaces can come to joint understanding and decisions concerning an environment. Both together form a testbed that could facilitate experimentation around social interaction and also provide a demonstration of this functionality in a remote collaborative simulation.

Most NVCs, such as interpersonal distance, interactional gaze and gestures of familiarity, are linked in social interaction. People obey these basic social rules when sharing a virtual environment with even a very simple virtual human [48]. However, this has not been tested with an avatar representation or with faithful reconstruction of identity and facial expression. This is not surprising as technology to support it has not been readily available. *Withyou* may be the first system able to support it. In addition eye gaze can be estimated to within the tolerances of social interaction from a video reconstructed avatar captured by cameras outside immediate social space [13] [44]. However, while support for each of these components had been tested, their linkage had not.

This linkage between these non-verbal behaviors is associated with familiarity and trust and is used to mediate interactions, deciding for instance if a conversation should start and when it should end. Because of this today's communication technology is generally less effective in anything but round the table meetings, where people's movement is constrained to a seat. This kind of meeting has been conducted using telepresence [22], however, bringing

someone's attention to something in the surroundings is hard when people cannot move around each other within a shared context. We argue that confusing or restricting spatiality of gaze and interpersonal distance holds back three applications of communication technology: the ad-hoc meeting; building of trust and rapport; and shared exploration of an environment, be it real, virtual or mixed. Supporting part of the CROSS DRIVE application by the integration of the Mars simulator has provided a case study and platform to test if these linkages are supported and ultimately if this makes a difference to task performance and experience. Fig. 13, demonstrates that a remote person can be viewed from any side without degradation of quality. This is unlike most many approaches that capture a person from predominantly one side [32][33].

Unfortunately, NVCs may lose their meaning if the shared space is not correctly orientated between participants. The linkup test showed a pragmatic approach to share the space between legacy display systems of distinct levels of immersion. In simple terms, displays that surround and immerse a user can be linked so that people can walk around each other. However, if any display does not surround the user, people can at most walk up to each other's avatars [44]. The legacy displays of the CROSS DRIVE partners were of both kinds. The pragmatic solution of projecting onto the floor in front of a wall display was used to provide a compromise. In this compromise, people can comfortably move within each other's social space without the need for those by a wall display to stand right up against it. Fig. 8, demonstrates this configuration, with one avatar just within and one outside natural social space. A seemingly "catch 22" problem is that 3D is needed to allow mutual eye gaze between moving people, yet 3D glasses or HMD occlude the eyes, Fig. 9. Our pragmatic solution to this is not to use stereo but rely instead on parallax to support mutual gaze. We had already shown that gaze could be accurately determined from a 3D model without stereo glasses [13] by participants rotating reconstructed humans until they appeared to look at them. Here we have tried it out for real, all be it not in a rigorous experiment.

Our approach provides a compromise between the qualities of video and VR. One of our goals is to balance visual, spatial and temporal qualities to the point where they can support the linkage between interpersonal distance, eye gaze and facial expression. The other goal is to provide an application that encourages people to move around and discus. Visual quality appears sufficient to communicate through the face: identity; gaze; and emotion, Fig. 7. The granularity of other non-verbal communication scales to clear finger gestures. In terms of spatial quality, our capture and reconstruction technique scales to full 360 degrees around the subject. While one of our display systems matches this, the others do not. Latency is much lower than what has been demonstrated by other full free viewpoint systems. However, it is still 1.5 seconds. This may impact on mutual eye gaze behaviour and role of non-verbal communication in conversational turn taking. Further improvements are being carried out by using a less conservative time management approach and we believe this will make possible to achieve latencies of around half this.

VI. Conclusions

We have presented the combination of the "withyou" telepresence system and a Mars simulator across visualization facilities in Germany and the UK. This pilot begins to demonstrate how space scientists, engineers and mission controllers could discuss Mars missions without the need to travel to single simulation facility in one of the participating countries. The contribution of this work is the combination of a telepresence system that allows spatial contextualization of an unprecedented range of Non-Verbal Communication (NVC), with a simulation that provides context and application that demonstrates utility, while addressing a set of key arising challenges.

Currently communication technologies do not well support the linkage between interpersonal distance, mutual gaze and facial expression. A key problem is that NVC is inherently spatial yet retaining spatial context across these non-verbal resources is hard without diminishing the quality of some. With our approach, gaze, interpersonal distance, facial expression and other NVC can be communicated as people move together around a place of interest, such as a landing site. This is challenging as previous telepresence systems have placed greater restrictions on different aspects of NVC, thus not allowing a natural set to be used together. The ultimate aim is to allow people to efficiently and accurately communicate both what they are talking about and how they feel about it, within the context of a shared simulation. Today, this is much easier when people are physically together. In widening the set of NVC that can be contextualized within simulations shared across a distance, this work could ultimately impact across many domains.

An important contribution is opening the door to technology mediated social interaction in which the linkage between interpersonal distance, mutual gaze and facial expression are likely to play a role. Both the photographs in this article and the supporting video provide evidence that people observe natural rules of interpersonal distance (e.g. Fig. 6) and gestures to bring them to it (e.g. Fig. 6 and Fig. 8). However, only one of the avatars was reconstructed from video. The others were conventional CGI avatars, two driven from motion tracking and another one from a desktop interface. This is the only system we are aware of that has both avatars created using video reconstruction and authored CGI models following motion tracked data. This might allow comparison of how the two approaches impact on the level of non-verbal communication supported during social interaction and outcomes such as trust, rapport, team cohesion and task performance.

A practical contribution is the demonstration of pragmatic approaches to share space between legacy display systems of distinct levels of immersion. For example, we showed how a wall display was extended with floor projection in order to improve the support for non-verbal behaviors in the social space of the users. The different displays allow the impact of display to be studied for the first time with video reconstructed

avatars.

A key novel technological contribution of this article is a new method for background segmentation. Segmentation allows a person to be captured without their surroundings so that they can be "imported" live into a shared virtual context. Previous methods supported segmentation against plain color or static backgrounds. However, the legacy displays used by the CROSS DRIVE partners had distinct levels of immersion. This meant that some people needed to be segmented from a moving CGI background, some from a static background, and others from a combination of the two. While the principle and technical feasibility has been demonstrated, a complete solution has not yet been implemented. This is simply as doing so requires the purchase of more of the equipment that we have already used. Specifically, the current implementation covers only part of our largest immersive display, needing the reminder to be turned off.

The rigor of this work is in the iterative steering of technology development from psychological principles. Other work has tended to focus on the support of subsets of non-verbal communication necessary for particular classes of interaction. We are not aware of work from other groups that has looked specifically at supporting the all-important linkage between interpersonal distance, mutual gaze and facial expression. Furthermore, we have addressed an unprecedented spread of issues to balance visual, temporal and spatial qualities. We further argue that this work will contribute to the rigor of future work by allowing us providing more ecologically valid social interaction experimentation.

Immediate impact includes demonstration to the space science community, how such technology could improve distributed team cohesion and reduce cost of international collaboration. This approach could be implemented in many other fields that require remote participants to discuss information or models that they need to move around together, especially where emotions are part of the conversation Joint emergency services command and control of a disaster scene is a good example of where both spatial context and strength of feeling need to be communicated together. Health applications could include remote exposure therapy and a better understanding of the importance of linked non-verbal cues, in interactions with virtual humans during training and self-treatment. However, the widest impact may come from adding knowledge to telepresence research, on the conditions that need to be met before the above linkage plays its proper role in starting and mediating conversations. Understanding this could lead to general rather than niche approaches to bringing people together across a distance. This could radically reduce dependency on travel and improve quality of life.

APPENDIX

Video footage of the linkup test can be found in the following URL: https://vimeo.com/141524309

ACKNOWLEDGMENT

We thank Johannes Hummel, Fang Chen, Wito Engelke and Andreas Gerndt from DLR, and John O'Hare from the University of Salford for their support in the work described in this paper.

REFERENCES

[1] D. J. Roberts, A. J. Fairchild, S. P. Campion, J. O'Hare, C. M. Moore, R. Aspin, *et al.*, "withyou—An Experimental End-to-End Telepresence System Using Video-Based Reconstruction," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 9, pp. 562-574, 2015.

[2] R. Westerteiger, A. Gerndt, and B. Hamann, "Spherical Terrain Rendering using the hierarchical HEALPix grid," 2011.

[3] A. Garcia, D. Roberts, T. Fernando, C. Bar, R. Wolff, J. Dodiya, *et al.*, "A collaborative workspace architecture for strengthening collaboration among space scientists," in *Aerospace* 2015.

[4] P. Milgram and F. Kishino, "A Taxonomy of Mixed Reality Visual-Displays," *Ieice Transactions on Information and Systems*, vol. E77d, pp. 1321-1329, Dec 1994.

[5] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, *et al.*, "Virtual space teleconferencing using a sea of cameras," in *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, 1994.

[6] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The office of the future: A unified approach to image-based modeling and spatially immersive displays," in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, 1998, pp. 179-188.

[7] M. L. Patterson, "An arousal model of interpersonal intimacy," *Psychological Review*, vol. 83, p. 235, 1976.

[8] V. Vinayagamoorthy, M. Garau, A. Steed, and M. Slater, "An eye gaze model for dyadic interaction in an immersive virtual environment: Practice and experience," in *Computer Graphics Forum*, 2004, pp. 1-11.

[9] D. Roberts, R. Wolff, O. Otto, and A. Steed, "Constructing a Gazebo: supporting teamwork in a tightly coupled, distributed task in virtual reality," *Presence: Teleoperators and Virtual Environments*, vol. 12, pp. 644-657, 2003.

[10] M. Slater, A. Rovira, R. Southern, D. Swapp, J. J. Zhang, C. Campbell, *et al.*, "Bystander responses to a violent incident in an immersive virtual environment," *PloS one*, vol. 8, p. e52766, 2013.

[11] Faceshift. (2015, 01/09/15). *Markerless facial motion tracking system [Online]*. Available: http://www.faceshift.com/

[12] Vicon. (2015, 01/09/15). *Vicon optical motion tracking system [Online]*. Available: http://www.vicon.com/

[13] D. J. Roberts, J. Rae, T. W. Duckworth, C. M. Moore, and R. Aspin, "Estimating the gaze of a virtuality human," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, pp. 681-690, 2013.

[14] M. Hansard, S. Lee, O. Choi, and R. P. Horaud, *Time-of-flight cameras: principles, methods and applications*: Springer Science & Business Media, 2012.

[15] P. M. Will and K. S. Pennington, "Grid coding: a preprocessing technique for robot and machine vision," presented at the Proceedings of the 2nd international joint conference on Artificial intelligence, London, England, 1971.

[16] G. C. Stockman, S. W. Chen, G. Hu, and N. Shrikhande, "Sensing and recognition of rigid objects using structured light," *Control Systems Magazine, IEEE*, vol. 8, pp. 14-22, 1988.

[17] A. Maimone and H. Fuchs, "Reducing interference between multiple structured light depth sensors using motion," in *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*, 2012, pp. 51-54.

[18] A. Maimone and H. Fuchs, "A First Look at a Telepresence System with Room-Sized Real-Time 3D Capture and Large Tracked Display," presented at the International Conference on Artificial Reality and Telexistence (ICAT), Osaka (Japan), 2011.

[19] A. Maimone and H. Fuchs, "Real-Time Volumetric 3D Capture of Room-Sized Scenes for Telepresence," presented at the Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), Zurich (Switzerland), 2012.

[20] A. Maimone and H. Fuchs, "Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras," presented at the Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on, 2011.
[21] D. S. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras," *Multimedia, IEEE Transactions on*, vol. 15, pp. 339-358, 2013.
[22] C. Zhang, Q. Cai, P. Chou, Z. Zhang, and R. Martin-Brualla, "Viewport: A distributed, immersive teleconferencing system with infrared dot pattern," *MultiMedia, IEEE*, vol. 20, pp. 17-27, 2013.
[23] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach," presented at the Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, 1996.
[24] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 1362-1376, 2010.
[25] B. G. Baumgart, "A polyhedron representation for computer vision," in *Proceedings of the May 19-22, 1975, national computer conference and exposition*, 1975, pp. 589-596.
[26] J.-S. Franco and E. Boyer, "Exact polyhedral visual hulls," in *British Machine Vision Conference (BMVC'03)*, 2003, pp. 329--338.
[27] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3d full human bodies using kinects," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, pp. 643-650, 2012.
[28] S.-Y. Lee, I.-J. Kim, S. C. Ahn, H. Ko, M.-T. Lim, and H.-G. Kim, "Real time 3D avatar for interactive mixed reality," in *Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*, 2004, pp. 75-80.
[29] R. Aspin and D. Roberts, "Projective multi-texturing for integrated real-time 3D reconstruction and rendering of a person," 2011.
[30] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. De Aguiar, N. Ahmed, *et al.*, "Floating textures," in *Computer Graphics Forum*, 2008, pp. 409-418.
[31] S. Beck, A. Kunert, A. Kulik, and B. Froehlich, "Immersive group-to-group telepresence," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, pp. 616-625, 2013.
[32] K. Higuchi, Y. Chen, P. A. Chou, Z. Zhang, and Z. Liu, "ImmerseBoard: Immersive Telepresence Experience using a Digital Whiteboard," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2383-2392.
[33] J. Zillner, C. Rhemann, S. Izadi, and M. Haller, "3D-board: a whole-body remote collaborative whiteboard," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 471-479.
[34] R. Vasudevan, G. Kurillo, E. Lobaton, T. Bernardin, O. Kreylos, R. Bajcsy, *et al.*, "High-quality visualization for geographically distributed 3-d teleimmersive applications," *Multimedia, IEEE Transactions on*, vol. 13, pp. 573-584, 2011.
[35] Y. J. Benezeth, P.-M.; Emile, B.; Laurent, H.; Rosenberger, C., "Review and evaluation of commonly-implemented background subtraction algorithms," presented at the Pattern Recognition, 2008.
[36] C. Schultz, "Digital Keying Methods," ed. University of Bremen Center for Computing Technologies, 2006.
[37] L. Li, W. Huang, I. Y. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 2-10.
[38] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 28-31.
[39] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters*, vol. 27, pp. 773-780, 2006.
[40] O. Grau, "Studio production system for dynamic 3D content," in *Visual Communications and Image Processing 2003*, 2003, pp. 80-89.
[41] J. O'Hare. (2015, 01/09/15). *Octave - technical information, University of Salford [Online]*. Available: http://www.salford.ac.uk/computing-science-engineering/facilities/octave-technical-information
[42] A. Griesser, S. De Roeck, A. Neubeck, and L. Van Gool, "GPU-Based Foreground-Background Segmentation using an Extended Colinearity Criterion," in *Proceedings of Vision, Modeling, and Visualization (VMV) 2005*, 2005, pp. 319-326.
[43] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: algorithms and evaluation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, pp. 918-923, 2003.
[44] T. Duckworth and D. J. Roberts, "Parallel processing for real-time 3D reconstruction from video streams," *Journal of Real-Time Image Processing*, vol. 9, pp. 427-445, 2014.
[45] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *Robotics and Automation, IEEE Journal of*, vol. 3, pp. 323-344, 1987.
[46] 7-ZIP. (2015, 01/09/15). *LZMA algorithm [Online]*.
[47] P. Ekman and D. Matsumoto, "Facial expression analysis," *Scholarpedia*, vol. 3, p. 4237, 2008.
[48] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis, "Equilibrium theory revisited: Mutual gaze and personal space in virtual environments," *Presence*, vol. 10, pp. 583-598, 2001.

Allen J. Fairchild is a PhD student at the University of Salford, under the supervision of David Roberts. His PhD seeks to refine an experimental 3D reconstruction telepresence platform to sufficient visual, spatial and temporal quality, to allow observers to link nonverbal cues across resources. He previously worked in PRimA developing pattern recognition software.

Simon P. Campion is a Project Manager and 3D generalist at the THINKLab, where he manages a portfolio of commercial Virtual Reality projects. He is also undertaking a part time PhD under the supervision of David Roberts. His PhD is assessing the impact of medium and interface on non-verbal communication.

Arturo S. García is a Research Fellow at the THINKLab. He received his European PhD in Computer Science at the University of Castilla-La Mancha (UCLM), Spain, 2010. His research interests include the development of Collaborative Virtual Environments (CVEs) and interaction in VEs and CVEs.

Robin Wolff leads the 3D Interaction group at Simulation and Software Technology at the German Aerospace Center (DLR), where he has worked as senior researcher since 2010. He received a PhD in Immersive Collaborative Virtual Environments at the University of Salford, UK, in 2007 and was working there in several research projects.

Terrence Fernando is the Director for the ThinkLab. He has a broad background in conducting multi-disciplinary research programmes involving large number of research teams in areas such as distributed virtual engineering, virtual building construction, driving simulations, virtual prototyping, urban simulation, and maintenance simulation. He was the technical director for the EU funded CoSpaces IP project which studied the challenges in creating a reference architecture that can support a range of collaborative environments. In this work he also investigated the challenges in developing co-located, distributed and mobile workspaces that can offer range of collaboration styles through innovative virtual interfaces.

David Roberts is a Professor of Telepresence at the university of Salford. He builds and studies the use of immersive VR, until now mostly for telepresence. c. 100 publications mostly in the area of telepresence or distributed simulation. The telepresence area of his work now focuses on video based reconstruction of humans in real time. He is part of the EU CROSS DRIVE project. David recently lead the EPSRC funded Eyecatching project which developed a telepresence approach that could support mutual eye gaze between moving people in different displays. He chaired IEEE Distributed Simulation and Real Time applications for 6 years.

Appendix C

Brining the client and therapist together in Virtual Reality Telepresence Exposure Therapy

D J Roberts[1], A J Fairchild[1], S Campion[1], A S Garcia[1], R Wolff[1]

[1]University of Salford
Salford, United Kingdom

[1]*www.salford.ac.uk/research/health-sciences/research-groups/virtual-reality*

ABSTRACT

We present a technology demonstrator of the potential utility of our telepresence approach to supporting tele-therapy in which client and therapist are immersed together. The aim is to demonstrate an approach in which a wide range of non-verbal communication between client and therapist can be contextualised within a shared simulation, even when the therapist is in the clinic and the client at home. The ultimate goal of the approach is to help the therapist to encourage the client to face a simulated threat while keeping them grounded in the safety of the present. The approach is to allow them to use non-verbal communication grounded in both the experience of the exposure and the current surroundings. While this is not new to exposure therapy, the challenge is to do this not only when the threat is simulated but when the client and therapist are apart. The technology approach combines immersive collaborative visualisation with free viewpoint 3D video based telepresence. The potential impact is to reduce dropout rate of exposure therapy for resistant clients.

1. INTRODUCTION

Exposure therapy appears to be the most effective treatment for phobias and post traumatic stress disorder (PTSD). Yet is suffers high dropout rates, especially in resistant populations, such as those with PTSD. One problem is that exposure therapy opens a wound to dress it, and opening it too quickly can retraumatise. Another is getting people with high levels of anxiety, and especially hyper vigilance, to reliably make it from the home to the clinic.

Virtual Reality Exposure Therapy (VRET) offers the potential of both heightened engagement and dosage control. Engagement helps with opening the wound and learning to close it. Dosage control could stop it being opened too much or too quickly.

Traditional exposure therapy, for example imaginal, uses different methods to manage a client's emotional distance to threat. A common approach is to maintain the client's grounding in the present, offering a safe container from which they can reach. This approach relies on the therapist observing non-verbal signals of the client and communicating with them, in part non-verbally, to bring their attention to the present. The typical approach of using Head Mounted Displays (HMD) in VRET and desktop displays with scaled CGI (Computer Graphics Imagery) virtual humans in lone and remote VRET poorly fits this. Our approach uses single wall immersive projection allowing the therapist to be visible along with the simulated threat. By using free viewpoint 3D video based reconstruction, our avatars communicate a far greater range of non-verbal signals than CGI avatars, including importantly, both attention and appearance. While not yet fully implemented, due to lack of equipment, our approach should allow the therapist to accurately determine if the client is looking at them, fixating on or away from the simulated threat, or following instructions to look toward the real world . We also demonstrate how video based reconstruction can be used to create 3D recordings of actors which have a visual appearance far more realistic than a traditional virtual human, and do so in a matter of minutes rather than weeks.

Proc. 11th Intl Conf. Disability, Virtual Reality & Associated Technologies
Los Angeles, California, USA, 20–22 Sept. 2016

2. RELATED WORK

VRET has been studied across an extensive range of phobias but perhaps most deeply with Post-Traumatic Stress Disorder (PTSD). Within PTSD, VRET has demonstrated potential efficacy and appears to be more engaging to resistant groups (Gonçalves et al., 2012). Yet drop-out rates, at approaching 40%, remain similar to non-technology based exposure therapy (Gonçalves et al., 2012).

Awareness of both memories and current present-moment experience is seen to facilitate exposure in traumatised individuals (Rothschild, 2003). Rothschild explains how the therapist uses non-verbal communication to detect fixation and bring attention back to the present. Conversely, "immersive virtual environments can break the deep everyday connection between where our senses tell us that we are and where we actually are located and whom we are with" (Sanchez-Vives and Slater, 2005). Most VRET uses Head Mounted Displays (HMD) (Gonçalves et al., 2012) that completely block both the present surroundings and therapist from view.

Tele-VRET has been demonstrated but uses desk-top interfaces through which tiny avatars representing client and therapist come together in a shrunken world. Such systems support little non-verbal communication or feeling of togetherness (Roberts et al., 2015a). People seem to react to life-sized virtual humans as if real, following natural patterns that relate gaze and interpersonal distance (Bailenson et al., 2001). Subtle changes in gaze and posture of virtual humans alters people's comfort (Pertaub et al., 2002). People respond naturally to virtual avatars in distributed collaboration between linked IPTs (Steed et al., 2005). We have extended such systems to support mutual eye-gaze (Roberts et al., 2009). However, these avatars still do not look like the person whose movements they copy and do not reproduce faithful facial expressions. We have thus developed 3D video telepresence to communicate both what someone looks like and what they are looking at (Roberts et al., 2015a). This technology produces live 3D graphical copies of people, and any items around them, into another space.

Others combined video based reconstruction with an immersive display (Gross et al., 2003) demonstrating how spatial and visual qualities could be better balanced. However, visual and temporal qualities were still some way behind what could be achieved with motion tracked avatars. Since then, visual qualities of video based reconstruction have significantly improved (Grau et al., 2007), (Waizenegger et al., 2011). Other recent (Divorra et al., 2010) and current (Steed et al., 2012), (Garcia et al., 2015) funded EU research focuses on spatial telepresence.

The potential utility of our approach in collaborative work has been demonstrated within the realm of space science and exploration (Garcia et al., 2015, Roberts et al., 2015b). This technology could ultimately be used to join clinic and home.

3. OUR TELEPRESENCE SYSTEM

The ultimate aim of our telepresence system is to situate people from different physical locations into a shared simulated context within which they communicate through a wide range of non-verbal resources. Unlike 2D video based approaches, each can see where the other is looking as they move. This system has been described before (Roberts 2013). Here we summarise what it tries to solve, its approach and current state.

Unlike spoken word, non-verbal behaviour and its use in social interaction is inherently spatial. Just as words link together to provide meaning, so do various non-verbal signals, along with their context. In the natural world, gaze, interpersonal distance and other non-verbal cues of familiarity are linked and used to allow people to manage relationships with each other. Even board room meetings typically importantly start and end with people going up to each other, making eye contact, smiling and sometimes tapping a shoulder or shaking hands. It is these things that grow trust between people, that is then useful in the meeting itself.

Video conferencing supports the kinds of interaction that we often have between those that help to develop trust and togetherness. Such technology ranges from Skype on a phone to carefully aligned screens and cameras around a table. Video allows the representation and linking of most non-verbal resources used in everyday conversations and interactions. However, it loses much of the spatial grounding. Spatial context can only be accurately determined within the space of the observed, rather than across the spaces of the interactants. While cameras and screens can be aligned to support some approximation of gaze interaction, this only begins to work when people remain stationary. Problems of aligning camera and image of face, and the Mona Lisa effect greatly limit this approach and make it completely unsuitable for supporting relationships between gaze and interpersonal distance. Video conferencing can be said to faithfully communicate visual but not spatial qualities of non-verbal behaviour.

Immersive Collaborative Virtual Environments (ICVE) offer the other extreme, where non-verbal communication between interactants can be situated in a shared virtual context but at the expense of visual faithfulness and many subtleties, such as facial expression. In such a system, people in different displays can move around a shared context together, seeing each other as life sized CGI avatars. We have previously extended ICVE with eye gaze (Roberts VR'09). Such a system theoretically supports the relationship between personal space and eye gaze although this has not been tested with rigour. ICVEs can be said to faithfully communicate spatial but not visual aspects of non-verbal behaviour.

Numerous video based approaches to reconstructing humans have been applied to telepresence. In theory, these should be able to faithfully communicate both visual and spatial qualities of non-verbal communication. However, balancing the two, especially with temporal qualities remains challenging (Roberts, 2013). This is the challenge that our telepresence system is set against. Specifically we want to faithfully communicate both visual and spatial aspects of non-verbal communication to within the limits of their use in non-touch interaction. This means being able to, for example, look someone in the eye and see if they smile as you enter their personal space, perhaps from the side.

Our approach combines real time free viewpoint video with immersive projection displays. An end to end description of the system is given in (Roberts et al., 2015a) . It adopts the video based construction approach of visual hull, using our parallel adaptation (Duckworth and Roberts, 2014) of the EPVH algorithm (Franco and Boyer, 2003) . Users stand within an immersive display system and are captured by surrounding cameras. Silhouettes from the images are then used to shape carve a form, onto which the original images are textured. Part of this process is shown in Figure 1. This live textured model can then be sent to another immersive display system to be placed within the spatial context of a shared simulation and another user.

We have built many prototype versions that between them demonstrate that all the fundamental requirements are achievable with our approach. However, we have not yet built a single version that fully meets all. At this point in time, we are able to build demonstrators of principle and undertake perceptual experimentation such as (Roberts et al., 2013). However, we currently lack the equipment and time to complete an end-to-end system that would demonstrate a sufficient balance of visual, spatial and temporal interaction to support meaningful behavioural experimentation. This paper presents a novel demonstrator.

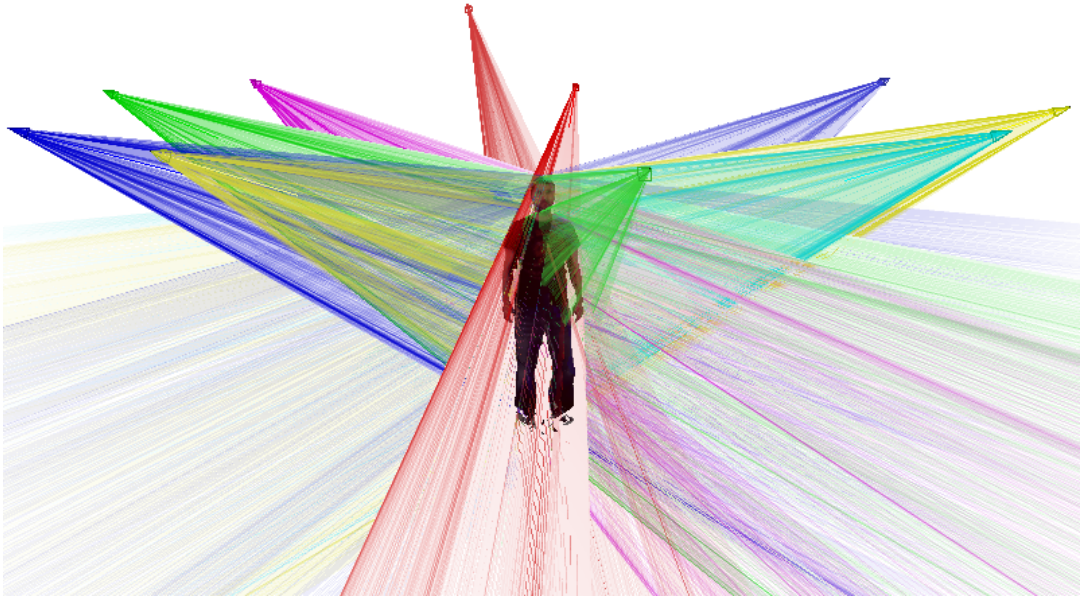


Figure 1. 3D reconstruction of a human in our telepresence system, while showing lines from each camera derived from silhouettes.

4. DEMONSTRATION OF THIS SYSTEM APPLIED TO VIRTUAL REALITY TELEPRESENCE EXPOSURE THERAPY

We begin by describing: the problem we are trying to solve; our general approach; an example scenario; the technology set up; and the limitations.

The problem that we are trying to solve is managing the emotional distance to threat while: 1) the threat is simulated; 2) the client and therapist are in different buildings. The approach we are taking is inspired by

Proc. 11th Intl Conf. Disability, Virtual Reality & Associated Technologies
Los Angeles, California, USA, 20–22 Sept. 2016

Rothschild (Rothschild, 2003) who attempts to mediate a client's awareness of threat and safety of the present, making use of verbal and non-verbal communication.

Our approach is to share a virtual context through large displays while using video based reconstruction to recreate both the therapist and, in this case, the threat. In another case the threat might be completely virtual. The concept is that the therapist can interpret both attention and emotion of the client through non-verbal signals and use non-verbal communication to bring their attention away or back to the threat to manage emotion.

In this scenario, the shared virtual environment represents a non-threatening place. The therapy scenario is one of social anxiety. The three people Figures 2 and 3 are authors playing out parts. One is the client, another the therapist and the other the threat. In Figure 2 the "client" looks straight at a threat that has just approach through a door. In Figure 3, the "therapist" steps between them and uses gesture and gaze to direct the client's attention to a neutral object, the sofa.

To demonstrate this principle and primary issues we have created an asymmetric system by linking two large displays with two different kinds of mediums (Figure 3). Asymmetric telepresence systems have been used to demonstrate the impact of differences in VR technology on collaboration (Slater et al., 2000) (Roberts et al., 2003). Our demonstration does not attempt to address every issue but does attempt to demonstrate the key issues and the fundamental qualities of our approach towards addressing them.

The novel configuration of our system used used in this demonstrator is shown in figures 4 and 5. Figure 4 gives a pictorial representation whereas figure 5 is shows system architecture. The client side uses very simple technology that would be relatively straight forward and inexpensive to replicate in the home. The key components are a large flat screen onto which the shared virtual context is displayed and a camera. A useful option is support for parallax so that the view into the other room moves with the viewpoint of the client. While this is not being used in the figures in this paper it was available and simply requires the "client" to wear markers, perhaps on his glasses or a hat. The therapist side is more complicated but could still be replicated within a clinic without excessive disruption or expense. The two fundamental differences are the use of two screens and a ring of cameras. The face on view of the "client" is transmitted via skype to a display wall in front of the "therapist". The rear portion of the partially shared virtual environment is displayed behind the therapist. A ring of cameras looks down at the therapist from the height of the screens. Each is angled so that while capturing the therapist moving within a portion of the space, neither screen is seen. This allows us to use a simpler and faster method of background segmentation that does not need to account for moving images. Between these two displays, the therapist can look ahead to see the client and behind to see the back of the virtual room the client looks into. The "therapist" appears in the foreground of the partially shared virtual space, as seen by the "client".

Figure 2. The "Client looks at an approaching virtual threat. The threat is a prerecorded 3D reconstruction of someone posing aggressively toward where the client now stands.

Figure 3. The "Therapist" moves between client and threat, and tries to redirect client attention.

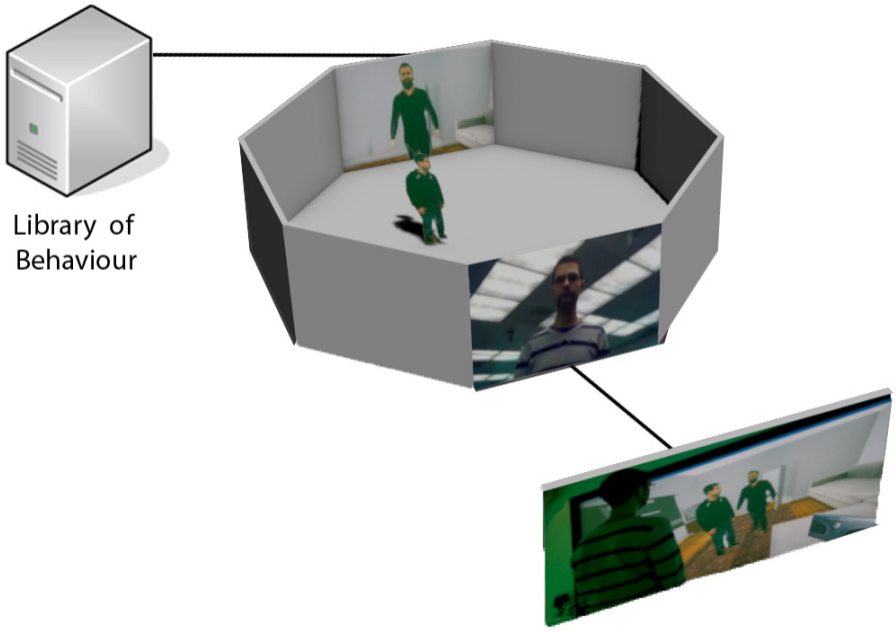


Figure 4. *Diagram of set up of an asymmetric telepresence system built for this demonstrator.*

Proc. 11th Intl Conf. Disability, Virtual Reality & Associated Technologies
Los Angeles, California, USA, 20–22 Sept. 2016

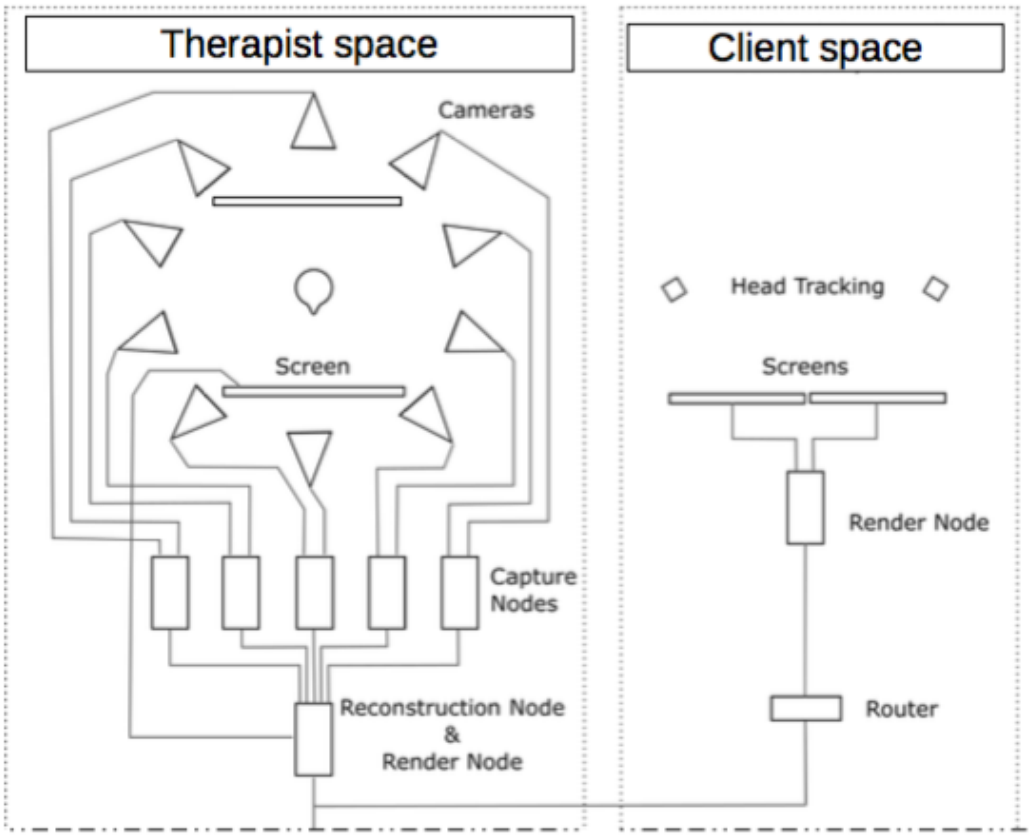


Figure 5. System Architecture

5. DISCUSSION

Our demonstrator is sufficient to demonstrate concept, approach and issues. It is however, currently incomplete, requiring more equipment to fully implement. Currently the lack of cameras at "client site" means the client cannot be reconstructed from video to appear within the therapist's physical space. Even if we had the cameras, we would still need to face the challenge of segmenting the user from a background displaying moving images and currently do not have sufficient infrared emitters to do this. We can represent the client in the therapist's space using a motion tracked avatar . Furthermore we can segment while one wall or one wall and the floor in front of it are displaying images. However, as we have demonstrated similar capabilities within a collaborative working application (paper about to appear) (Fairchild et al., 2015), we decided not to provide new evidence here.

While our implementation is in the above ways incomplete, it does demonstrate an approach that with more equipment would meet full requirements. The equipment used does actually help to demonstrate the issues we attempt to solve and the utility of the solution. Using a single wall rather than surround projection for the client has a number of advantages. Most importantly, it allows them to stand in and still see the real world. It is also more deployable and less obtrusive within a home setting. A commodity projector focussed on a matt whitewall would suffice. This does, however, mean that the therapist cannot enter their space. We have demonstrated how such a set up can allow the therapist to move between the client and virtual threat. We demonstrate in a paper about to appear (Fairchild et al., 2015) how stereo glasses and a second projector aimed at the floor can allow an avatar to enter local space. We could at present project onto the surface in front of and under the feet of the therapist, while still being able to subject this image to reconstruct their form and likeness. We have chosen in this early paper to simply use 2D video (Skype) to stream the images of the client to the therapist. When cameras are carefully aligned, this allows the therapist to see if the client is looking at them, past them at the simulation, or into the client's own physical space. While this might allow the therapist to step into the line of sight of the client, it is important to remember they would no longer be able to determine the client's gaze from that position. In order to maintain the ability to estimate the client's gaze, we would need a camera ring around the client and to reconstruct them in 3D. Furthermore, while our segmentation approach should work against most static backgrounds, this has not been rigorously tested.

6. CONCLUSION

We have demonstrated the principle of supporting a wide range of non-verbal communication contextualised within a shared virtual reality therapy environment. We have explained how this could be applied to help a therapist keep their client grounded in the real world as they confront a simulation of what they fear.

This is not the first time that immersive projection technology has been used in VRET. However, we are unaware of a publication describing its use to support non-verbal togetherness of client and therapist or communication between them. This is not the first time that immersion and life sized avatars have been used to improve feelings of togetherness or contextualise non-verbal communication. For example, we have previously described our technology approach to faithfully communicating both appearance and attention by combining immersive displays with free viewpoint 3D video based avatars. We have also previously described its application to collaborative working. This is the first time its potential application to exposure therapy has been described.

The demonstrator we have presented is meant to convey approach and issues rather than be a complete and fine-tuned system. It demonstrates a range of technologies put together in a pragmatic way. Both simpler and more advanced approaches could be derived from this. The most advanced would allow a full sharing of virtual context in which client and therapist could move around together. The current and simpler versions provide a partial sharing of context and impose restrictions on movement within the space. However, the demonstrated and simpler approaches are far more deployable and affordable given current technology.

The potential impact of this approach is in reducing dropout rates of exposure therapy. This is important as dropout rates of 40% are not uncommon in resistant populations. Furthermore, as symptoms typically increase at the beginning of a course of exposure therapy, clients can dropout with negative health impacts. We argue that by allowing clients to both use virtual reality exposure therapy and work with a therapist at home reduces the risk of non-attendance to therapy sessions. This could impact not only on success rate of treatment but in reducing costs to health providers through reducing missed appointments. We further argue that allowing the therapist and client to see each other and estimate what the other is looking at, would help to manage the grounding of the client in the safety of the present. This again has potential to reduce dropout rates by reducing the risk of retraumatisation and improving the relationship between client and therapist. While remote therapy can be done with conventional video conferencing and CGI avatars, the levels of non-verbal communication used within a clinical therapy session are not supported. Our approach has the fundamental properties to support them much better. Our demonstrator shows both the issues and the principles of the solution.

Acknowledgements

The authors with to thank Charlie Moritz from Freedom form Torture, Allan Barret from Pennine NHS Care Trust, Warren Mansell from University of Manchester and Linda Durbrow-Marshal from University of Salford for helping us understand the relationship and interaction between client and therapist and what needs to change in VRET to accommodate this. We also wish to thank the technology team at Salford that have helped in the past to develop the telepresence system. This includes Toby Duckworth, Carl Moore and Rob Aspin and John O'hare who manages the oCtAVE and helps with everything done within it.

7. REFERENCES

Bailenson, J N, Blascovich, J, Beall, A C & Loomis, J M (2001). Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence,* 10**,** 583-598.

Divorra, O, Civit, J, Zuo, F, Belt, H, Feldmann, I, Chreer, O, Yellin, E, Ijsselsteijn, W, Van Eijk, R & Espinola, D (2010). Towards 3D-aware telepresence: Working on technologies behind the scene. *Proc. ACM CSCW: New Frontiers in Telepresence*.

Duckworth, T & Roberts, D J (2014). Parallel processing for real-time 3D reconstruction from video streams. *Journal of Real-Time Image Processing,* 9**,** 427-445.

Fairchild, A J, Campion, S P, Garcia, A, Wolff, R, Fernando, T & Roberts, D J ---2015. A Mixed Reality Telepresence System for Collaborative Space Operation.

Franco, J-S & Boyer, E. (2003) Published. Exact polyhedral visual hulls. British Machine Vision Conference (BMVC'03). 329--338.

Garcia, A, Roberts, D, Fernando, T, Bar, C, Wolff, R, Dodiya, J, Engelke, W & Gerndt, A. (2015) Published. A collaborative workspace architecture for strengthening collaboration among space scientists. Aerospace Institute of Electrical and Electronics Engineers.

Gonçalves, R, Pedrozo, A L, Coutinho, E S F, Figueira, I & Ventura, P (2012). Efficacy of virtual reality exposure therapy in the treatment of PTSD: a systematic review. *PloS one,* 7**,** e48469.
Grau, O, Hilton, A, Kilner, J, Miller, G, Sargeant, T & Starck, J (2007). A free-viewpoint video system for visualization of sport scenes. *Motion Imaging Journal, SMPTE,* 116**,** 213-219.
Gross, M, Würmlin, S, Naef, M, Lamboray, E, Spagno, C, Kunz, A, Koller-Meier, E, Svoboda, T, Van Gool, L & Lang, S. (2003) Published. blue-c: a spatially immersive display and 3D video portal for telepresence. ACM Transactions on Graphics (TOG). ACM, 819-827.
Pertaub, D-P, Slater, M & Barker, C (2002). An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoperators and virtual environments,* 11**,** 68-78.
Roberts, D, Wolff, R, Otto, O & Steed, A (2003). Constructing a Gazebo: supporting teamwork in a tightly coupled, distributed task in virtual reality. *Presence: Teleoperators and Virtual Environments,* 12**,** 644-657.
Roberts, D, Wolff, R, Rae, J, Steed, A, Aspin, R, Mcintyre, M, Pena, A, Oyekoya, O & Steptoe, W. (2009) Published. Communicating eye-gaze across a distance: Comparing an eye-gaze enabled immersive collaborative virtual environment, aligned video conferencing, and being together. Virtual reality conference, 2009. VR 2009. IEEE. IEEE, 135-142.
Roberts, D J, Fairchild, A J, Campion, S P, O'hare, J, Moore, C M, Aspin, R, Duckworth, T, Gasparello, P & Tecchia, F (2015a). withyou—An Experimental End-to-End Telepresence System Using Video-Based Reconstruction. *Selected Topics in Signal Processing, IEEE Journal of,* 9**,** 562-574.
Roberts, D J, Garcia, A S, Dodiya, J, Wolff, R, Fairchild, A J & Fernando, T. (2015b) Published. Collaborative telepresence workspaces for space operation and science. Virtual Reality (VR), 2015 IEEE. IEEE, 275-276.
Roberts, D J, Rae, J, Duckworth, T W, Moore, C M & Aspin, R (2013). Estimating the gaze of a virtuality human. *Visualization and Computer Graphics, IEEE Transactions on,* 19**,** 681-690.
Rothschild, B (2003). *The body remembers casebook: Unifying methods and models in the treatment of trauma and PTSD*, WW Norton & Company.
Sanchez-Vives, M V & Slater, M (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience,* 6**,** 332-339.
Slater, M, Sadagic, A, Usoh, M & Schroeder, R (2000). Small-group behavior in a virtual and real environment: A comparative study. *Presence,* 9**,** 37-51.
Steed, A, Roberts, D, Schroeder, R & Heldal, I. (2005) Published. Interaction between users of immersion projection technology systems. HCI International 2005, the 11th International Conference on Human Computer Interaction. 22-27.
Steed, A, Tecchia, F, Bergamasco, M, Slater, M, Steptoe, W, Oyekoya, W, Pece, F, Weyrich, T, Kautz, J & Friedman, D (2012). Beaming: an asymmetric telepresence system. *IEEE computer graphics and applications***,** 10-17.
Waizenegger, W, Feldmann, I & Schreer, O. (2011) Published. Real-time patch sweeping for high-quality depth estimation in 3D video conferencing applications. IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 78710E-78710E-10.